THE FORM AND AUDITORY CONTROL OF DOWNWARD TRENDS IN INTONATION

Michael Edward Johnson

submitted for the degree of PhD

University College London, September 1993

To
Mum and Dad,
Sylvia and Peter
and
Martine
sine quis non

ABSTRACT

Of all the areas of intonational research, study of the tendency of the frequency of vocal fold vibration to decline during the course of an utterance - F0 declination - is likely initially to be the most fruitful in determining the interaction between perceptual and productive processes. A general introduction to the phenomenon is augmented by analysis of different methods of determining declination lines; theoretical treatments are then introduced. One particular local factor contributing to the downward trend, downstep, is discussed, and its pivotal role in the intonational phonology developed by Janet Pierrehumbert critically examined. In the light of the theoretical discussion, two competing hypotheses are presented as to the mediation of the declination effect, which is the effect that of two accented syllables in an utterance, the second has to have a lower peak F0 value than the first for them to be judged to have equal prominence. The Global Declination Hypothesis attributes this to the use by speakers and hearers of one or two abstract reference lines declining through the course of a tone-unit. The Local Declination Hypothesis attributes it to the disposition of F0 excursions surrounding the two accents as well as to the respective peak values.

The Global Declination Hypothesis is tested by presenting listeners with pairs of dual-peak accented utterances with the two peaks identical in F0, without any physically present local declination, and asking them to rate the prominence of the second peak of each such utterance. No significant differences are found in the prominence ratings, so the Local Declination Hypothesis appears to be favoured. That hypothesis is itself tested through the development of a model of individual accent prominence, which incorporates terms for surrounding unaccented context. This is then used as the basis of a model of the perceptual constraints on the production of intonation in the scaling of target peaks. The model predicts that local slope between accents and slope of the context after the target accent, as well as other local variables, jointly determine the F0 value of a peak with a particular targetted prominence relationship with its predecessor. If the interaccentual stretch is declining, the declination effect is predicted to occur, ceteris paribus. The model is found to be initially acceptable. In addition, a global interpretation of downstep is made within the model.

The mechanisms the model is suggested to represent are auditory feedback control loops of a variety of possible degrees of complexity. An experiment is devised to test for the basic existence of a feedback loop which is used to prevent local slope exceeding an arbitrary threshold value. Auditory feedback in subjects was disrupted by headphone-administration of low-pass filtered masking noise during their utterance of a sustained vowel, and a short and a long dual peak-accented sentence. The disruption was sufficient to alter the apparent mechanism controlling the production of the sustained vowel, but the Lombard effect, whereby subjects automatically raise the level of their voice in ambient noise, was found to be a vitiating factor.

General conclusions are drawn on the nature of the declination phenomenon in intonation, and proposals made for future research.

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Prof. A.J. Fourcin, for his sympathetic encouragement. I have been motivated and encouraged similarly by other members of the Department of Phonetics and Linguistics at UCL, notably Mark Huckvale.

During the short time I spent on study leave (on the ERASMUS scheme) at the University of Nijmegen in the Netherlands, I was greatly assisted by many members of the Phonetics Institute there. I would like to express my gratitude to them, and in particular to my supervisor there, Dr. Toni Rietveld.

The research was funded by the Medical Research Council, whose assistance I gratefully acknowledge.

Deepest acknowledgment is to Martine Grice, for health, hearth and heart.

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

1.1 INTRODUCTION

The study of intonation, as the study of any phonetic phenomenon, can be performed from a number of different angles: auditory, productive physiological, acoustic, perceptual physiological; these are some of them. The length of the phonetician's task would extend to many lifetimes if he or she had to perform detailed research into intonation in each one of these areas. However, at the same time as comprehensive work is impossible by one individual in the archetypically multi-disciplinary study of phonetics, so it is necessary that the place of a particular line of research within the multi-disciplinary framework is continually assessed. At the same time, it is dangerous to ignore the possibility that a particular line of research might coordinate two distinct research areas. There is one area of intonation research in which the need for a coordinated research programme is either essential, or unnecessary, depending on your viewpoint, and that is the study of fundamental frequency declination. In the most general terms, this is the phenomenon whereby a downward trend in the course of the fundamental frequency of vocal fold vibration is displayed over an utterance.

There are those who view it as an inevitable accompaniment to speech, because it is attributable wholly to passive phenomena resulting from the physiological constraints of the speech production mechanism, notably subglottal pressure decline as a result of on-going lung volume reduction (e.g. 't Hart et al, 1990), and relaxation of vocal fold tensor muscles, particularly the cricothyroid muscle, (e.g. Titze and Durham, 1987). This point of view holds that declination can be studied from a purely physiological point of view, that it can be treated as a global phenomenon, albeit with a phrasing function, and otherwise factored out of intonation contours being studied.

An alternative point of view is that it is the very fact that the form of declination appears to be constrained by productive physiological mechanisms which makes it a natural object of coordinated study - particularly of research into both productive and perceptual mechanisms involved in the generation of intonation trends. This is because, firstly, and trivially, apparently safe hypotheses ought to be challenged, and secondly, for normal

speakers, production and perception go hand in hand; and there is no better object of study for investigating how those processes cooperate on a day-to-day basis, and how they develop ontogenetically and have developed phylogenetically, than F0 declination. The very first utterance each of us makes has a simple declining unaccented F0 contour. It is the initial exercise we perform as a basis for later speech, and the most rudimentary. It is therefore a natural starting point for an investigation into the coordination of productive and perceptual processes in intonation. This alternative viewpoint is the one propounded in this thesis.

In the next section, the more general object of study, intonation, is discussed. More formal reasons for studying F0 declination are then introduced. Then the quantity whose variation is to be studied is discussed, and finally, an outline is presented of the thesis.

1.2 OBJECT OF STUDY

In this section are discussed the experimental methods which can be used to test out hypotheses regarding the productive and perceptual correlates of intonation contours produced by speakers of a language. The nature of these mechanisms will depend to a large extent on the very hypotheses that are used to initiate the research. Any linguistic hypothesis (that is, any hypothesis 'about language', where language is to be interpreted as a phenomenon, without presupposing its natural constituency[1]) must include reference to system primitives or parameters, independently of the linguistic function which manipulates those parameters. As Scuffil (1982) writes, in his consideration of how intonational data should be regularised,

"A major theoretical and practical problem for the observer is the distillation of his explicanda from his data; in other words, what criteria he has for regularising his data." (Ch. 5, p.154).

At the same time, a linguistic hypothesis could include reference to the operation of the system, in other words, the processes that are used to manipulate the primitives that are identified as being retrievable from the raw

[1] Cf. Chomsky (1988) : ".. there is nothing in the real world corresponding to language. In fact it could very well turn out that there is no intelligible notion of language." (p.107).

data.

It is possible to ignore the system operation in intonational analysis: this is typically done in analyses which aim to account for intonation as part of a human speaker's linguistic competence. In these cases, the object of study is the intonational knowledge that a speaker has internalised as a member of a linguistic community. This knowledge is expressed as part of a fundamentally logico-mathematical model of linguistic capability called a generative grammar. Such a model also treats of operations and operanda, but these are, a fortiori, different animals from those that can form part of a directly testable hypothesis (Chomsky 1965).

The position adopted here is that an experimental investigation into intonation must take into account both the operation and the operanda of the intonation system as exhibited by human speakers in an experimental situation. The initial hypotheses will then form part of a tentative model of the intonational performance of a speaker of a language; that is, they will make claims about what is happening in the intonational part of the productive and perceptual apparatus of a member of a linguistic community when he or she talks or listens. Given the disjunctive nature of this model, any initial hypotheses must be framed in terms of intonational analytic parameters of both a productive and perceptual nature. It would be possible to postulate intonational performance parameters which logically are independent of experimentally determined lower-level productive and perceptual variables, but it would be wrong to do so without postulating how they interact with the lower level, because they are not directly retrievable in an experimental situation. It would be even more wrong to perform experimental intonational analysis with parameters which properly form part of a model of linguistic competence; such an analysis requires the relationship between competence and performance models to have been mapped out for it to be considered rigorous.

1.3 TYPES OF INTONATIONAL ANALYSIS

Analysis of intonation contours takes place at a number of different levels of human endeavour. It can be said that intonational analysis is performed by the brain of humans when they listen to someone talking their own language. In fact, the type of analysis performed in these circumstances varies

according to the degree to which listeners have acquired an intonation system for their language, or whether indeed the listener, if he or she be an infant, has developed an intonation system at all for a particular language.

Intonational analysis of a sort is also performed by someone listening to an unfamiliar language. The intonation patterns presented to such listeners would be to some extent alien, depending on how closely the intonation system of that language matched their own (although certain forms could be interpreted in ways that conform to their own intonational conventions - which, inter alia, might result in a misinterpretation of the attitude of the speaker). The type of analysis performed in these circumstances would be similar to the type of analysis performed by a phonetician when transcribing the intonation contours of a particular language. In ideal circumstances, such analysis would be performed without any linguistic preconceptions, such that in the first instance, it would correspond to a purely auditory analysis (although as the phonetician were to develop his system of analysis, the categorisation that resulted would tend to damage the purity of the auditory record).

However, as Ashby (1990) points out in his general discussion of phonetic categorisation, such ideal circumstances cannot seriously be considered ever to obtain in the field. Instead, the phonetician categorises sound patterns according to a set of prototypes, the clearest example of which is the set of cardinal vowels in the system developed by Daniel Jones. Sounds are said to more or less closely approximate to particular prototype sounds :

"Our introspections, as we work out our response in relation to some stretch of speech tend to be along the lines of 'that's some kind of a [t]', or 'that's a vowel something like [i]'"
(Ashby 1990:23)

This is in contrast to the type of analysis which is attributed to a standard view of the phonetician's work, within which categorisation is performed according to Aristotelian principles : categories are determined by intersecting sets of independent attributes. Such categories are of a kind with those of distinctive feature theory (Jakobson, Fant and Halle, 1952, Chomsky and Halle, 1968).

The analysis that an auditory phonetician performs is thus of a particular, distinctive kind, and can only be referred to with the undecomposable denotation 'phonetic'. It is on the boundary between what we might call natural human speech analysis (since phoneticians are human) and linguistic theoretic speech analysis, in which the categories identified are taken to form part of a model of speech processing capability. Within the realm of intonation, a great number of phonetic analyses have been presented in published form. (It is worth mentioning here that the Atlantic ocean splits this body of intonation analysis in two, such that traditional British treatments of (English) intonation employ prototypes which are pitch configurations such as 'fall', 'level' and 'fall-rise', and traditional American treatments prototypes which are pitch levels. This means that the classical British phonetician identifies intonation patterns in relation to an inventory of prototypical pitch movements, while the classical American phonetician identifies intonation patterns in relation to a restricted set of turning points of pitch movements within that inventory. In both cases, however, the approaches have formed part of a pedagogical tradition of intonational analysis. Thus, many systems of intonation which derive from phonetic analysis are directed towards teaching purposes, with some resultant necessary simplification.

On the linguistic theoretic side of intonational study are phonological analyses of intonation, which can be divided into classical taxonomic phonemic or tonemic analyses and the more modern generative analyses, which themselves can be grouped into linear and non-linear varieties. These terms will be used in this thesis, where appropriate, in relation to individual analyses; for the time being, it may be said that modern phonological analyses of intonation are (or should be) part of the description of a putative speaker's intonational competence, in the sense referred to above. They treat of a set of tonal categories and their manipulation, either within a segmental string by sets of reorganisational phonological rules (the linear approach), or within domains distinct from that of the surface segmental string (the non-linear approach). These analyses are thus somewhat abstract, in the sense outlined by Cutler and Ladd (1983:1ff). Classical tonemic analyses stem from a more direct structural grouping of intonation patterns into formally similar categories which, because they have been distilled by a systematic approach of phonetic commutation, are considered to be recognised and used by speakers of a particular language, and thus

suitable for communication to learners of that language in a pedagogical context.

Although dealing with abstract entities, the methods adopted by generative grammarians are similar in principle[2] to those of communication engineers who analyse speech by

(a) developing models of speech production,

(b) using them to synthesise speech, and

(c) comparing the synthesised speech, within a particular context, with natural speech, the closeness of fit being a mark of the accuracy of the model (cf. Flanagan, 1972). The similarity is particularly marked in the case of terminal analog models of speech production, in which it is only the synthetic speech signal which is expected to match its human counterpart, the production processes being non-comparable (Holmes et al, 1964); a generative grammar is similarly tested against data, but typically the dataset is that formed by intuitive evocation by an analyst. These methods are sufficient only to satisfy the criterion of observational adequacy of the grammar (Chomsky 1965), but there are problems even satisfying this criterion when it comes to the intonational component, because intuitive evocation of intonational data is not so robust as that, say, of syntactic forms (although the grammaticality of the latter can be ambiguous in some cases). In particular, the different ways of partitioning the intonation contour, and the detail allowed in each intuitive analysis, allow for considerable variation in the inventory of pitch contours which a model has to match.

As a result, and also simply because there is a current trend for acoustic-phonetic corroboration of phonological theory (see Ohala and Jaeger 1986), some generative analyses of intonation require the criterion of observational adequacy to be satisfied in respect of hard data in the form of F0 traces. The result is a set of synthetic models of intonation which incorporate rules developed within the generative paradigm (e.g. Pierrehumbert, 1981). These are distinct from models of intonation which simply adopt engineering principles of analysis to synthesise intonation contours (e.g. Edward, 1982). The types of intonational analysis which incorporate the methodology of analysis-by-synthesis could be referred to as synthetic analyses of

---

[2]Chomsky claims the methods to be quite different in fact; generative grammars are not just engineered systems; they are hypotheses about the real world.

intonation. Those that specifically incorporate models of intonation developed within the generative paradigm could be referred to as generative-synthetic.

At a lower and more experimental level of intonational study are analyses of the productive physiological concomitants of intonation. These analyses treat of the low-level determinants of intonation contours, particularly laryngeal and pulmonary activity. Intonation contours can then be modelled in terms of the interaction of relevant production processes. The purely passive physical and aerodynamic aspects of intonation, specifically in respect of vocal fold vibration, are equally important areas of study. The work by Titze (e.g. Titze, 1973, 1974, 1989, Titze and Talkin, 1979) is a notable contribution here, as are the biennial symposia on Vocal Fold Physiology (Stevens et al, 1981, Bless and Abbs, 1983, Titze and Scherer, 1983, Baer et al, 1987, Fujimura, 1988, Gauffin and Hammarberg, 1991).
Analyses can also be performed of auditory function in respect of pitch perception; such perceptual physiological treatments are equally as partial in respect of the complete intonation system as their productive low-level counterparts.

Explicitly productive and perceptual models of intonation tend not to incorporate directly such low-level concomitants. Instead, higher level control or perceptual functions are used as the building blocks for the intonation contour. In productive models, the control functions are largely hypothetical entities viewed as being implicated in the coordination of groups of muscular activity[3]. In perceptual models, the perceptual functions are usually determined by psychoacoustic or psychophonetic experimentation - they may correspond, for example, to particular primitive types of pitch movement[4].

1.3.1 Summary of types of intonational analysis

It will be clear from the above that the many forms of intonational analysis differ in their generality. The first type mentioned, natural human

---

[3] e.g. Ohman 1967. Fujisaki 1987 is an exception to this rule, in that he proposes a model of intonation in which the phrase component and accent component are equated with independent activity of separate parts of the cricothyroid muscle (pars obliqua and pars recta).

[4] see Collier 1989 for a summary of the Dutch approach.

intonational analysis, is clearly the object of study, but phonetic intonational analysis is of course also performed by humans, and the knowledge of intonation gained by this means is extremely valuable. (Nonetheless, this should not prevent the intonational judgments of phoneticians being themselves objects of study; indeed such an enterprise can be valuable for intonational research). In respect of this generality, the experimental types of intonational analysis mentioned, viz. productive physiological, aerodynamic, audiological and psycho-acoustic can only offer partial accounts of how human beings process intonation; they don't say enough of detail about the intonational capacity of a human speaker-hearer, and how it manifests itself in performance. Each of the other types of scientific analysis must also leave many things unsaid; the point is clear that a successful unified theory of intonation will result only from the consideration of all the types of analysis listed, along with their interaction with analyses of pragmatic function.

In the latter regard, it should be underlined that the orientation of all the types of analysis referred to is in some sense formal. Intonational function is largely ignored, although many auditory phonetic analyses try to map out the difficult relationship between intonational form and function (Halliday, 1967, Crystal 1969).

The types of intonational analysis can be summarised under the following headings :

Object of study	Type of theoretical/experimental intonational analysis
Natural human intonation analysis.	Auditory Phonetic Pedagogical Phonological: Phonemic/tonemic Generative: Linear Non-Linear Synthetic: Generative-synthetic

Otherwise
Productive physiological
Aerodynamic
Perceptual physiological
Psycho-Acoustic

1.4 A SIMPLIFICATION OF THE ANALYTICAL POSITION

The list of types of analysis at the end of the previous section could be augmented by subdividing major headings, particularly those of the theoretical analyses. A thorough approach to the study of intonation would involve a cross-classification of all treatments of intonation to date in respect of their separate categories. The aim of this would be to discover exactly what each has to contribute to a unified theory of intonation, by studying carefully the implications of each treatment. Often this would involve addressing particularly thorny issues, such as how a phonetician's auditory/phonetic analysis can be compared with quantitative studies; that is, what the phonetician's auditory transduction function is, and how consistent it is from phonetician to phonetician; and, if phonetic analyses cannot hope each to be interpreted in respect of a common standard, to what extent statistical analysis might aid in distilling common intonational features from them which could be compared with quantitative data.

However, notwithstanding the difficulty involved in preliminary collation and cross-classification of extant treatments of intonation, the question of what the object of study consists of is itself a marked problem in the development of a unified intonational theory. John Ohala, who himself has worked to put intonational study onto a rigorous footing, writes :

"Phonetics has not, in fact, succeeded in explaining all of the interesting sound patterns involving intonation and tone. It has, however, made sufficient progress in explaining a subset of these patterns that we are justified to expect that continued research on physics, physiology, and perception of intonation and tone will eventually give us the answers we seek." (Ohala 1982, p.8)

which might be seen to imply that it has been possible unproblematically to ask relevant questions. Yet in 1970, he wrote :

".. speech research does not seem to be able to make all of its disparate pieces of data fit together into a general pattern of any sort. Typically it is unable to generalise beyond the avilable data and cannot tell what kind of experiments are worth doing." (Ohala 1970, p.1).

It is possible that in the intervening years, he has discovered enough to satisfy him that we are asking the right questions, but, for the current author, the problem of determining the right questions to ask is still pressing.

This is because there appear to remain certain fundamental problems in intonational analysis. Firstly, there is the problem of identifying the abstract tonal categories of the linguist (in particular of the generative variety) with productive and perceptual phenomena determined by experimentation. This stems from the fact that the account of the linguist still fundamentally relies on auditory/phonetic data, despite the appearance of compliance with objective acoustic reality in certain cases[5]. This is not to doubt that such data has validity. Some of the most incisive work on English intonation has been performed by two auditory analysts, Dwight Bolinger and David Crystal[6]. Yet their intonational constructs shouldn't have a pivotal place in a unified theory of intonation, because they are each mediated by a phonetician's perceptual model, which, as noted in section 1.3, is more than the auditory model of the naive listener. Like experimenters in Quantum Physics, the activity of phoneticians alters the conditions in the system they are trying to describe, but here because that system is a fragile construct within their own minds. The phonetician's model can only act as a <u>basis</u> for enquiring after the productive and perceptual correlates of intonation; its status as a hypothesis is even open to doubt if, following Popper[7], we require that hypotheses be formulated such that they be refutable by the

---

[5] e.g. Pierrehumbert 1980, who rejects auditory analysis, but tacitly relies on it at some points (notably in her discussion of stressed syllables in low tails) and certainly provides no formal mechanism for aligning her tonal categories with her F0 traces.

[6] See, for reference, Bolinger 1986 and Crystal 1969.

[7] e.g. Popper (1957)

adduction of a consistent body of counterexamples discoverable just beyond the state of knowledge that existed when the hypothesis was formulated. The hypothetical system of the phonetician can only be corroborated or refuted by the construction of an adequate model of intonation which can then be applied to the phonetician himself, along with some model of phonetic model construction.

Secondly, although a model of tonal production might be constructed which satisfactorily accounted for production of an intonation contour in any context, from initiation in the motor cortex (or even lower down in the laryngeal motor pathway) down to vocal fold activity; and although a model of tonal perception accounting for tonal transduction in the inner ear to interpretation in the auditory cortex might be constructed, still the following questions might be outstanding :

(i) How do tonal productive phenomena map onto tonal perceptual phenomena?
(ii) How are tonal phenomena organised (syntagmatically) within an intonation system; or more fully, how is the intonational system organised within the brain?

That is, is the productive-perceptual mapping via the mediation of more abstract tonal entities, or is there a more gradual merging of productive tonality into perceptual tonality, and vice-versa such that there are links between the two systems at lower than neocortical levels? In short, the essential problem is one of what constitutes a correct performance model of intonation.

Finally, many of the questions that are asked in intonational phonology are not directly relevant to experimental studies in intonation and vice versa because, as mentioned above, they are couched in generative terms such that it is presupposed they are relevant only to a model of intonational competence. Knowing what are the right questions to ask presupposes knowing what side of the competence-performance divide one is investigating.

Thus, in considering these problems, it seems that in aiming for a unified theory of intonation (which is, the author thinks, the proper goal of intonational research), the following <u>might</u> be the right questions to ask :

Q1. What, in neurophysiological terms, is the nature of the productive mechanisms by which intonation contours are generated?

Q2. How does the intonation system use those productive mechanisms?[8]

Q3. What, in neurophysiological and psycho-acoustic terms, is the nature of the perceptual mechanisms by which intonation contours are received and processed by the brain?

Q4. How does the intonation system use those perceptual mechanisms?

Q5. Within the Central Nervous System, and given its particular anatomy, what kind of neural network could act as the substrate to intonation system?

Clearly, these questions are firmly on the performance side of the competence-performance divide. Yet it is important to pursue at the same time phonological research into intonation, to the extent that it continues to produce descriptions in new and elegant terms of intonation systems[9]. Therefore, a more theoretical question which must remain addressable in the background of research is :

Q6. What is the relationship between a competence model and a performance model of language; specifically, of intonation?

This question does beg the question of whether the intonation system is properly part of a competence model of language. If it could be shown that no part of the intonation system is exclusively linguistic, then a performance model of intonation would be adequate to characterise the intonation system of a (group of) speaker(s). However, some languages at least use intonation

---

[8] By this question is meant, for instance, is each component of the tonal productive mechanism devoted to one particular function (e.g. Cricothyroid muscle for raising pitch, Posterior Crico-Arytenoid muscle for abducting the vocal folds), or can they be deployed in different ways to produce the same tonal effect ?

[9] What it, along with all generative-grammatical study to date, has failed to do, is provide an explanation for the phenomena it describes - Cf. Chomsky 1988 pp.27ff, i.e. how the system forms part of a communicating and, a fortiori, developing organism.

in part purely for grammatical contrasts - for example, English, where tonal form distinguishes the syntactic phrase marker for

She didn't go because \/Jack was there.

from that for

She didn't go because \Jack was there.

(the scope of 'not'; in the former sentence is the subordinate clause and in the latter is the main verb). Thus, since syntactic phenomena are uncontroversially linguistic phenomena, a linguistic analysis of intonation is very probably highly relevant to an understanding of intonation systems. In this case, a relevant question is

Q7. Where is the division in an intonation system between linguistic and non-linguistic uses of the phenomenon?

These seven questions cover three domains of analysis : productive (Qs 1, 2, 6), perceptual (Qs 3, 4, 6) and linguistic (Qs 6, 7). For each of these three domains, a different set of analytical parameters could be determined. An account of what these might be, and how they are interrelated, is one of the motivating aims of this thesis (see chapter 5 for consideration of some of the questions in the productive and perceptual domains). Because of the size of the task, a further reduction in the scope of the analysis has to be performed. This is done as a result of considering the general question whether the intonation system of a speaker could be independent of the physiological mechanisms through which it operates.

There are two extremes to this position: in the one, it could be said that the intonation system is completely dependent on the physiology of the speech mechanism, right down to laryngeal innervation and muscular control, and cochlear innervation and control; in the other, it could be said that it is a completely contingent matter that it uses these mechanisms - it could just as easily have used the tactile sensory modality, say, and relied on caresses and thumps for the communication of messages. Neither of these extreme views could seriously be taken, of course; a more reasonable position holds that the intonation system is dependent on specific human physiological mechanisms, but not fully so; one important question to be answered (in the fullness of time) is where in the speech chain that dependency begins.

It is clearly not the case that the intonation system requires a particular

detailed laryngeal strategy to be implemented for each contour type. Indeed, there are speakers who implement no laryngeal strategy whatever to produce intonation contours, viz. laryngectomees, who use electrical prostheses or oesophageal speech for this purpose; and even normal speakers can mimic intonation contours by the use of overall amplitude and spectral variation when producing whispered speech (though see footnote 10). Nonetheless, it remains generally true that normal speakers use normal phonation to generate intonation contours[10], and as Wiktor Jassem points out (in criticism of Hjelmslev's (1953) cenematic theory of language)

"The fact that many people can live without their legs does not necessarily mean that the very nature of human life on earth is that a man can do without his legs." (Jassem 1952:16).

For such normal speakers, it is likely to be the case that habitual productive mechanisms are employed for particular pitch patterns. It seems certain that the intonation system is more dependent on the auditory system. Without cochlear and auditory nerve stimulation of some sort, no intonation patterns can be perceived by an individual.

Thus, an initial hypothesis regarding the independence of the intonation system could be that it is more independent of productive mechanisms than it is of perceptual mechanisms, so that the production strategy of a speaker might be

"produce a tonal form that is unambiguously like tonal sound-image x, using any of the productive mechanisms available in the current context, but favouring those most commonly used".

This view of an intonational strategy would be similar to that proposed for articulatory phonetics in Ladefoged (1971).

One of the aims of this thesis is to see whether there are grounds for such

---

[10] It is also worth noting that at least one study of accented devoiced and voiced vowels in Japanese (Sugito and Hirose 1988) has noted that Cricothyroid muscle activity, which, as noted below, is highly correlated with upward pitch excursion, is associated in much the same manner for the devoiced forms (which are fully whispered) as for the voiced forms. The implication is that certain laryngeal muscular programs for accentuation may remain the same regardless of vocal fold configuration.

a hypothesis. Now, in general, the intonation system of an arbitrary language (irrespective of the existence of lexical tone in that language) can be characterised as comprising a phrasal component (i.e. a component which cues phrasing) and an accentual component (i.e. a component which highlights particular structures the sizes of which vary depending on the scope of accentuation) (cf. Fujisaki, 1987, 't Hart et al, 1990 Thorsen, 1980, Hirst, 1988).

As a strategy for reducing the scope of the research, a choice could be made between studying either one of these components. The one to be chosen would be that most likely to bear fruit in respect of the hypotheses about the role of perception in intonation production. Often, the phrasal component is manifested as the declining trend in fundamental frequency referred to as 'declination'. Since this phenomenon is often attributed to purely passive phenomena in the speech production mechanism, it makes sense to test for the existence of auditory targets in its generation. For if there is evidence that such targets exist in this component of intonation, which appears so unaffected by perceptual constraints, then the evidence could be considered to be heightened of the use of such perceptual targets in intonation in general (the current author does not necessarily adhere to this hypothesis).

Since evidence for the existence of of auditory targets in the production of declining trends requires some statement of what the form of those auditory targets should be, this thesis acquires the title it has[11]. It will be suggested towards the end of the thesis that statements of the form of declination properly belong to one particular field of study, viz. phonology.

1.5 THE QUANTITY USED TO EXPRESS THE VARIATION IN INTONATION

In experimental research, intonation contours tend to be expressed using the quantity fundamental frequency (F0). Strictly speaking, this quantity is only valid of a stationary periodic signal; its estimate is therefore always heuristic (see Hess, 1983). Most frame-based algorithms used for computing this quantity (i.e. those which are not period by period) use 30-40ms time

---

[11] The word 'declination' is not used in the title, as that term has come to be used for just one of few identified contributions to downward intonation trends, the most important of which is downstep. This latter phenomenon is discussed in Chaper 3, with a different view of it suggested in Chapter 5.

windows. This works because that duration approximates the lower limit on perception of pitch; such algorithms thus have some direct correspondence with putative processes of pitch perception.

Other algorithms detect a different quantity, related to the production of voice. Notable amongst these are those based on the laryngograph (see Fourcin,1974, Fourcin and Abberton, 1971) which uses the transmission of an AC current at a microwave frequency across the larynx via externally applied electrodes so that varying glottal impedance caused by vocal fold vibration may be recorded. The signal thereby produced is Lx, from which, by one of a number of algorithms derived to detect the point of maximal vocal fold contact (the point of closure) the quantity Tx is derived, which is a measure of local period of vocal fold vibration. The reciprocal of this quantity is Fx, corresponding to the fundamental frequency of excitation of the vocal tract.

The filterable signal corresponding to this quantity (the Fx signal) is bound to be modulated in some way during the transmission of speech both by bone conduction (cf. Tonndorf 1984) by which the excitation signal is transmitted directly to the ear of speakers as they are talking, and by air conduction; this passage includes the filtering activity of the vocal tract and the effects of radiation from the lips and transmission through the ambient environment to a hearer (which may be that same speaker). However, the relative energy in the signal is so high (compared with the energy in other frequency bands within the speech signal) that essentially the same quantity can be used for quantifying the signal received at the ear and used for the transmission of intonation patterns as used to quantify the signal emitted from the larynx.

Once the highly non-linear neurophysiological modulation of the auditory pathway comes into the picture[12], one might expect the quantities corresponding to perception and production to be more disparate. However, models based on psychophysical experiments for stationary tones do not display large differences in the quantity 'pitch' from the fundamental frequencies of the input stimuli. However, most pitch perception models don't

[12] For physiological models of auditory function at different levels of the auditory system, see, for instance, Meddis and Hewitt 1991a, 1991b, Hewitt et al. 1992, Hewitt and Meddis 1993 and Hewitt and Meddis (forthcoming).

approach the question of dynamic stimuli, and it could in theory be that the very quantity used to quantify the signal corresponding to the percept in the perception of intonation would better be scaled somewhat differently from F0 on the linear Hz scale. Hermes and van Gestel (1991) find that ERB-rate[13] is a more appropriate quantity for scaling intonation than either Hz or the semitone scale.

It is ultimately important to get the respective quantities right for production and perception in intonation, as some behaviour in their interaction could turn out to be accounted for simply by using the correct quantities. At the beginning of the research for this thesis, it was thought that an established model of pitch perception should be used as the basis for quantifying perceived pitch contours hypothesized to correspond to Fx contour stimuli. As a consequence, a dynamic form of Terhardt's algorithm (Terhardt et al. 1982) for the computation of virtual pitch (which is based on his analytical model of pitch perception - see, e.g. Terhardt 1974), was encoded on a minicomputer. The algorithm allows for the computation of two quantities - Nominal Virtual Pitch, which is computed from spectral pitch according to a summing of subharmonics determined by iterative frequency shift of the harmonics estimated from the power spectrum; and true virtual pitch, which corresponds more to the perceived pitch returned during psychophysical experiments, and incorporates pitch shifts of the order of 4% or so in a signal within the range of intonation. An example of the operation of the algorithm appears in Figure 1.1 . Virtual Pitch (VP) has been computed for a frame at the peak of the first accent in Figure 1.2, which shows the output VP contour (to which some postprocessing has been applied) superimposed on the Fx contour corresponding to the speech signal in the figure. The processes involved in computing the virtual pitch value are shown from top to bottom of the left column, then top to bottom of the right column. These stages correspond to the stages depicted in Terhardt et al. 1982. p.681. The speech in the first window has been windowed at 40ms using a Hanning window.

[13] See chapter 2 for the formula that they use for this quantity.

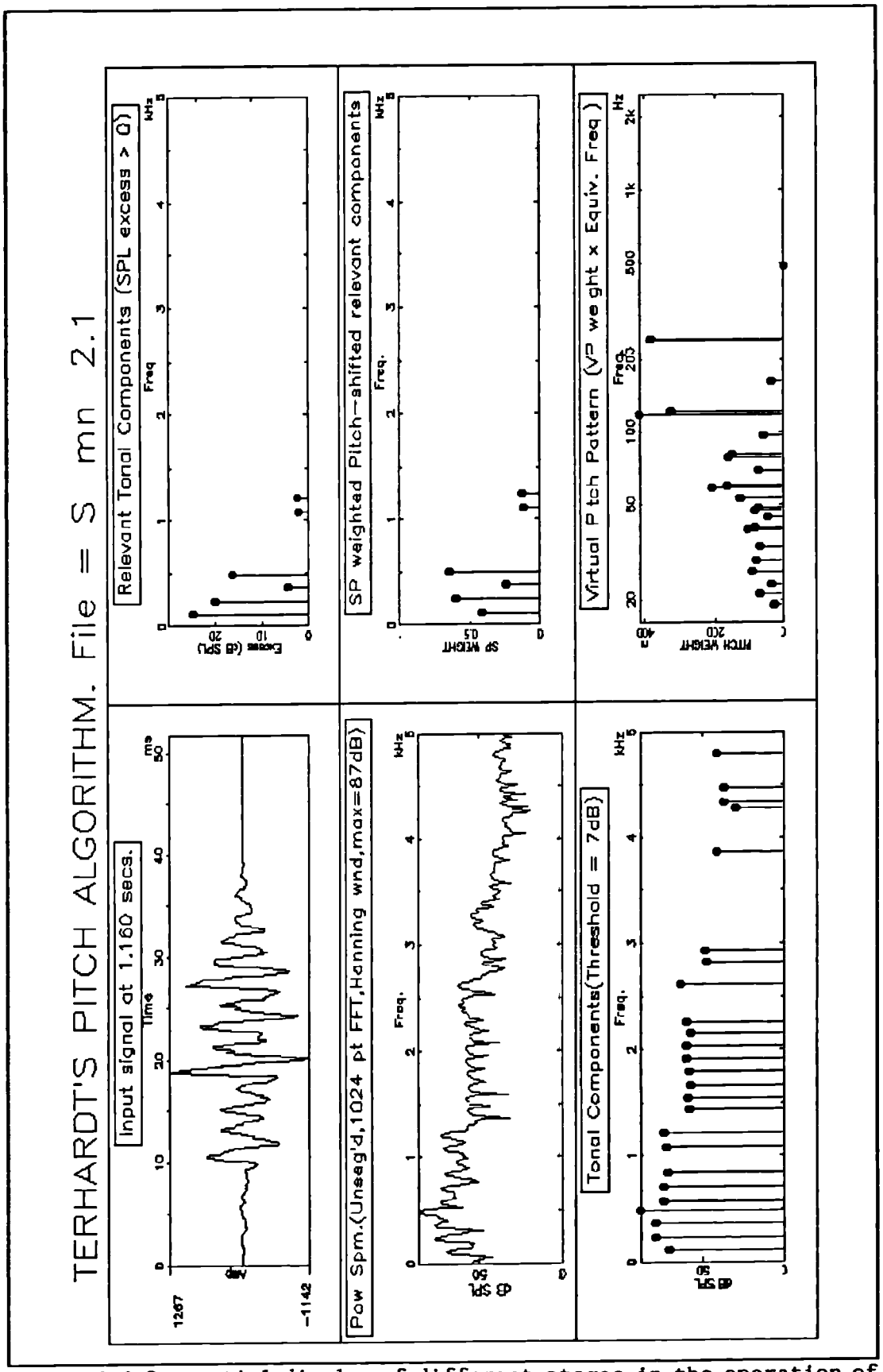


Figure 1.1 Sequential display of different stages in the operation of a dynamic version of Terhardt's model of pitch perception.

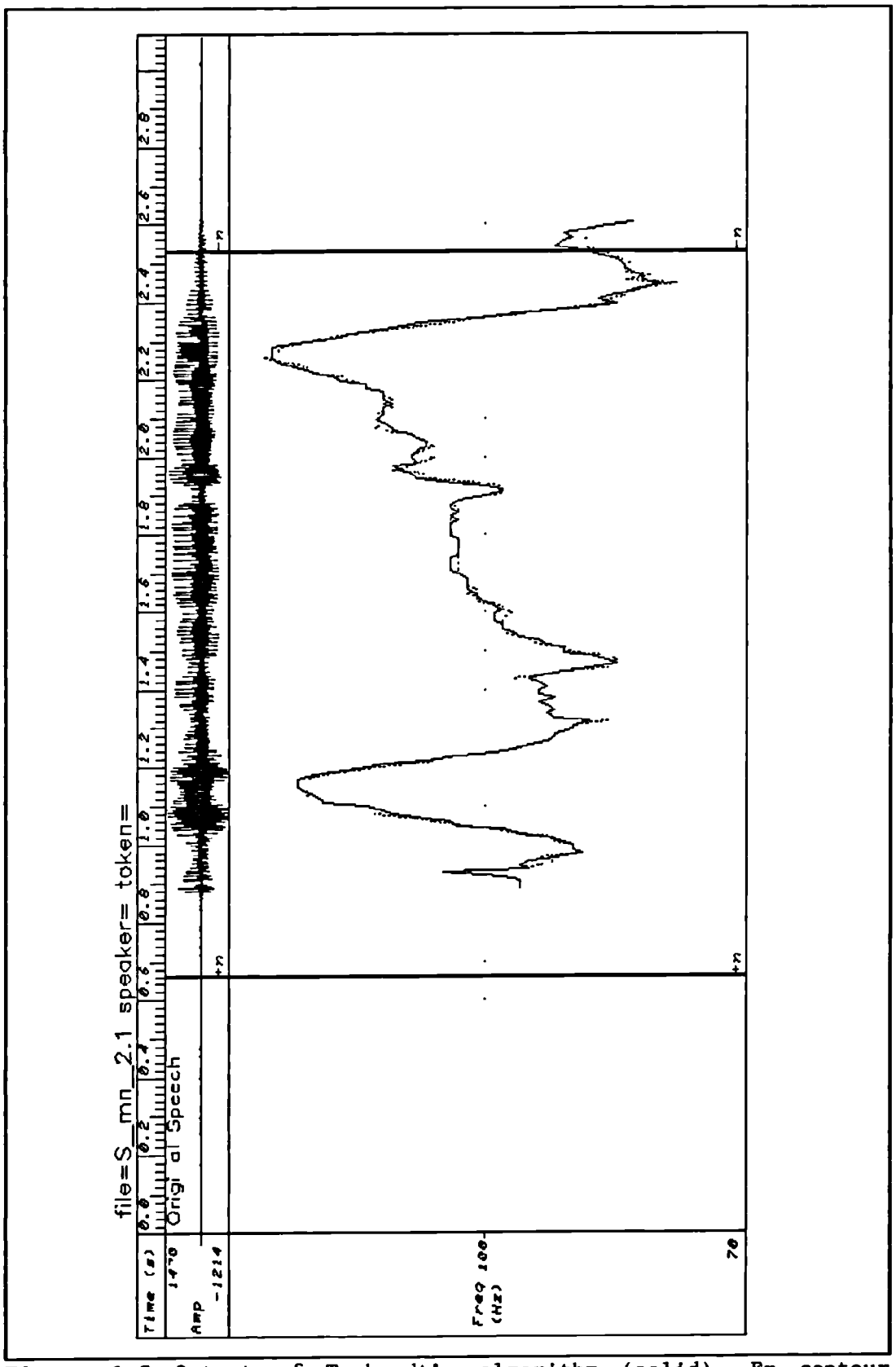


Figure 1.2 Output of Terhardt's algorithm (solid), Fx contour (dotted) and corresponding speech signal. The computed quantity is Nominal Virtual Pitch. Labels relate to experiment in Chapter 6.

In the event, the presuppositions involved in the model were found to be too unwieldy to incorporate within an open-ended investigation into the processes mediating the phenomenon of declination, so the quantity virtual pitch was not used to quantify perceptual counterparts of the intonation contours. Instead, the quantity F0 was used for the purposes of the investigation (and Fx in Chapter 6), and its expression in Hz. is found to be of utility in Chapter 5. However, Terhardt's algorithm was used in quantifying the intonation contours in the data of Chapter 6, when Fx contours were not available due to experimental error.

1.6 OUTLINE OF THESIS

The rest of this thesis is developed in five strands: a descriptive one, in which the phenomena of declination and downstep are described; a phonological one, notably in Chapter 3, in which a critical evaluation of Pierrehumbert's model of intonation is made; a meta-theoretical one, in which different theoretical approaches to declination are discussed; a computational modelling one, in which a model of pitch prominence is developed and used as a basis for modelling perceptual control of the production of downward trends in intonation; and an experimental one, in which some initial steps are taken in assessing the existence and nature of such control.

Chapter 2 gives examples of the phenomenon of declination, discusses standard partitions of it, and discusses different approaches to identifying declination lines, and more generally to modelling declination. It is observed that most treatments have declination as a global construct.

Chapter 3 concentrates on the phenomenon of downstep, and how it is used as the basis for Pierrehumbert's 1980 phonological model of intonation. Ladd's 1992 model of downstep is also discussed. The pivotal nature of the process to intonation is questioned, and a more restricted role for the phenomenon suggested.

Chapter 4 discusses experiments designed by other researchers to investigate the context-sentitivity of the declination effect, in which later peaks of equal prominence actually have lower peak F0 values. From consideration of the experiments, two competing hypotheses are proposed to account for the declination effect, the Local Declination Hypothesis and the Global Declination Hypothesis. The Global Declination Hypothesis is tested

by means of a perceptual experiment.

Chapter 5 investigates the validity of the other hypothesis, the Local Declination Hypothesis, by developing a model of perceptual determinants of the production of intonation directly from a model of pitch prominence based entirely on the configurations of salient points identifiable in a F0 contour. Downstep sequences are seen to be generable by the model by a more global production mechanism. The implications for the form and auditory control of declination are discussed, and difficulties in interpretation of the notion of form assessed. Two basic mechanisms for the production of intonation, one involving finer control in auditory feedback than the other, are proposed.

Chapter 6 presents a first experiment in the sequence required to investigate the putative mechanisms introduced in Chapter 5. In this experiment, just the existence of auditory feedback in the control of both the unaccented context of accents, and accents themselves, is tested using noise masking of auditory feedback. Results are considered in the light of the suggested production mechanisms, and of the interfering factor of the Lombard effect.

Chapter 7 summarises the thesis, speculates on neuro-physiological mechanisms which could possibly be involved in the control of F0 declination, and makes some suggestions for future research.

CHAPTER 2
THE PHENOMENON OF DECLINATION

2.1 INTRODUCTION

Having established the quantity whose variation is to be considered in this inquiry, it is time to look at the specific type of variation of interest here, that is, declination. After the discussion at the end of Chapter 1, our objects of study could be considered to be both declination in Fx and declination in virtual pitch, but in the rest of the thesis the more general view is taken for the purpose of introducing the phenomenon, and it is F0 declination that will be studied in some detail[1].

In addition, another, related phenomenon, which has been referred to as the 'declination effect', will be considered. This is the effect in many languages that a later accented syllable in an utterance will be considered to have the same prominence as an earlier one if (and possibly only if) its peak F0 is lower than the earlier one by a certain factor. This phenomenon is clearly closely connected to that of F0 declination; what is of importance is to discover how directly the declination effect depends on F0 declination. Later in the thesis it will be shown that investigating the connection in a particular way permits exploration of the relationship between perceptual and productive aspects of intonation.

2.1.1 The domain of declination

Fx is a direct function of vocal fold vibration, and varies from voice period to voice period. Thus, an increase in duration in one voice period over its predecessor is a reduction in Fx. However, it is clear that a reduction in Fx over one period cannot be considered to be an example of minimal Fx declination. So what does constitute such a minimal reduction? Similarly, pitch is a quantity that can be estimated over varying sizes of time-window, although there is a certain size of time-window below which it is not appropriate to talk of pitch perception of speech at all. Again, the reduction in pitch from one minimal perceptual frame to the next should not be

---

[1] Here we consider not declination of amplitude, or of loudness, which pair of quantities certainly undergoes a gradual reduction in many natural utterances, nor 'declination', if it can be called that, of articulatory precision, which with its perceptual counterpart, is also a common phenomenon in speech (c.f. Vayra and Fowler, 1992).

considered an example of declination, but what should? These questions can be reduced to the single one of what the domain of declination is, but they implicate another, viz. that of how declination is to be determined from the variation in the frequency quantity, here F0.

2.1.2 Methods of determining declination lines

To answer the latter question, we need first to consider how previous analysts have approached the phenomenon of declination. For most, the quantity from whose variation they determine declination is short-time averaged fundamental frequency or F0, rather than vocal-fold vibration frequency or virtual pitch, but this doesn't matter for the point at issue; for the time being, we can speak of F0 declination, on the understanding that the quantity F0 is measurable or estimable in some physical domain. In many studies, F0 declination is identified with a straight line, and sometimes a set of straight lines, which pass through the local maxima in an F0 contour, either in the linear or logarithmic frequency domain. Some studies identify an additional line which passes through the local minima. In the former case, the phenomenon of declination is called 'topline declination' and in the latter case, 'baseline declination'[2].

These studies fit the straight lines through the peaks and troughs according to different criteria, which can comprise a set of rules satisfying constraints which are either formal (in terms of the shape of the F0 contour) or perceptual (in terms of a matching auditory percept). There are other, less common studies which fit straight lines according to more general algorithms, for instance, those comprising the class of methods for determining a line of best fit through a set of data points by regression. These could either fit a

---

[2] It is immediately apparent that these treatments consider declination to be a partly abstract phenomenon, since
(a) either line usually only touches certain parts of the contour
and
(b) in parts of the F0 contour in which either line touches a contour whose F0 is declining, considered over a non-minimal time-window, local variation around the declination line is ignored or smoothed out.

In respect of the processing of intonation, these forms of abstraction can be considered to be on distinct levels, and this position will be assumed in later chapters.

line through the whole F0 contour, or through the peaks and troughs of the contour.

The method of linear regression, by itself, is the first stage in determining non-stationary trends from a sequence of non-independent data x_i (where the value of x_i typically can be shown to depend on the value of x_{i-1}) i.e. a time series (see, for example, Pandit and Wu, 1983). A second stage would involve examining the residuals from the regression to identify a second-level trend. The examination from regression of residuals has been used in a more indirect fashion to determine such a second-level (more restricted) declination function. Pierrehumbert and Beckman (1988) adopt the method of determining the declination line from the mean residuals calculated by sentence-length category from linear regression through pairs of consecutive F0 peaks in a number of related sentences. The assumption here is that the covariation of the respective peak values is accounted for by some function other than declination - for instance downstep (for which see Chapter 3) - with the possible augmentation of local prominence scaling. If the mean residual is greater in shorter utterances than in longer, it means that the difference between the accent peaks is greater than accounted for by, say, the downstep function, by a greater amount in shorter than longer utterances. On the assumption that the residual variance can be wholly attributed to a declination function, it implies a linear one of variable drop across utterances of different durations.

This underlines two important questions about declination models: (i) whether there is only one downtrend in any intonation contour, which could be referred to as 'declination', or whether declination is one of a number of such trends; and (ii) whether the declination line has constant drop[3], constant slope, or variable drop and slope. The first of these questions will be introduced in more detail later in this chapter.

A constant drop linear declination function has also been proposed (by Pierrehumbert 1980) as a more abstract component of a model of prominence scaling. In this case the declination line is determined by fitting the

[3] 'drop' refers to the F0 distance between the start and end of the declination line.

parameters of the model to the F0 contour. The resulting declination line need in principle have no stretch in common with the F0 contour. This raises the question of the degree of abstractness in the declination function (already referred to in fn. 1 above). That issue will adopt some importance in this thesis.

Other possible methods of computing declination lines include the estimation of a trend from the time series which constitutes the F0 contour by a method other than linear regression; that method typically would be an appropriate low-pass filter. This could never recover a declining linear trend perfectly; the closer an appropriate declining linear trend were retrieved, the more the method would be akin to determining a declining DC component from the F0 contour treated as a signal. The parallels with such an operation will be discussed later in this chapter, in connection with the issue of declination and componentiality in the F0 contour. The following section will look first at the other general method of determining declination lines, which in principle does not assume a prior interpretation of the intonation contour in terms of productive physiology or mechanics, psychophysics or phonology. It will then go on to consider the more specific methods which do take into account some knowledge about the intonation contour, be it expressed in terms of heuristic rules about the surface form of the contour or as a more abstract model of intonational form.

2.2 METHODS FOR DETERMINING DECLINATION LINES

2.2.1 Linear Regression of F0 on time

Perhaps the simplest method of computing a declination line for a F0 contour is by fitting a Linear Regression line through it. This constitutes the reduction of modelling the downward trend in an F0 contour to the simplest form possible, a linear model with one random variable (F0) and one fixed variable (time), with two parameters, an initial offset (= the intercept) and a rate of decline (= the slope). It is the approach taken by Lieberman et al.(1985) in a highly controversial paper which throws into relief the difficulties in determining a consistent heuristic strategy for the identification of reference lines in measured intonation contours.

In that paper, the aim was to find support for Lieberman's breath-group theory of intonation - in particular, that intonation contours can be divided into two main types - falling or rising - depending on what happens at the end of them. He and his associates place declination theory in opposition to this, implying that a declination theory will maintain the existence of a downward trend in all (spontaneous and read) utterances by which phrasal domains are determined by speaker-hearers. They chose linear regression as a method for computing declination lines after finding that Maeda's (1976) eye-fitting algorithm gave inconsistent results when applied to a sentence corpus. Further, they chose an all-points regression line rather than a regression line through heuristically determined peaks and troughs of contours.

Their main results were that declination slopes for spontaneous speech are far more variable than those for read speech, and that this is even more the case when terminal contours (the last 150 ms) are excised before computing the declination lines. In read speech, the variability of slopes was about the same in full and non-terminal (in which the terminal contour was excised) utterances. Summarised in this way, the results are not particularly conclusive.

The paper was heavily criticised (see Repp 1985) for statistical ineptitude in concluding, according to a strange criterion of contour linearity established by the significance of the correlation coefficient, that all-points linear regression lines (many degrees of freedom) were better choices than peak or trough points regression lines (few degrees of freedom). It was also criticised ('t Hart 1986) for failing to distinguish, amongst the sentences analysed for declination slope, between those having a late final fall, and those not. To this can be added the point that to use the correlation coefficient as a criterion of success in fitting a declination line implies that contours with a certain degree of disruption due to accentuation or segmental coarticulation effect will tend to exhibit less declination than those without, which there is no reason to believe is the case[4] - the slope of the regression line is a better estimate of the degree of declination in these terms. Apart

[4] Although there is some justification for thinking that accentuation excludes declination - see Chapter 4.

from these methodological criticisms, however, the following more theoretical ones can be levelled: firstly, Lieberman et al. assume that declination theories require a downward trend to be physically present in the F0 contour; secondly, the whole of a downward trend must be subsumed under one factor, declination, rather than being partitioned into separate factors, one of which, final lowering[5], can be equated with the falling terminal contours of Lieberman's breath-group theory.

Those issues will be discussed in due course. For the time being, it is appropriate to examine separately the method of linear regression for computing declination lines; some other problems inherent to the method will arise.

Figure 2.1 shows the F0 trace of a single tone-unit utterance with a regression-line fit by the least-squares method. What appears to be a reasonable declination line according to one sort of heuristic strategy, that it follow or parallel the F0 contour in stretches of unaccented speech, appears to have been achieved in this case.

Figure 2.2 shows another such trace, this time of a question with predominantly rising intonation, where the regression line fit is clearly not declining. The understanding on adoption of this method can be seen immediately to be that declination is a phenomenon observable only in some utterances. In fact, with the observation that declination lines computed by this method tend to have steeper and more consistent slopes in prepared (read) speech than spontaneous speech (Lieberman et al. op.cit., Umeda 1982) it would seem that declination has, if anything, a degenerate linguistic function, and cannot be used for phrasing because of its inconsistent ontological status.

For defenders of declination, such a conclusion is the death-knell for linear regression as a method for determining declination lines. However, even without any theoretical preconceptions, the method appears to apply too straightforward a model to what, in an F0 contour, is, despite the fact that

---

[5] The phenomenon of final lowering is examined in section 2.2.3.1 below.

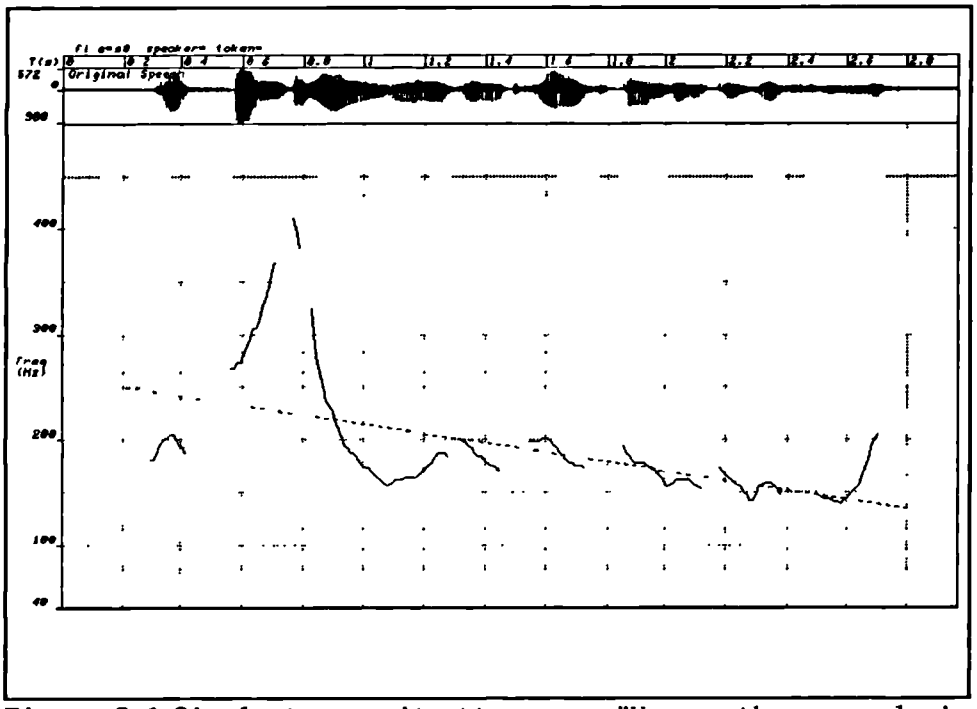


Figure 2.1 Single tone-unit utterance: "We sum the numerals in corresponding positions", Speaker F1

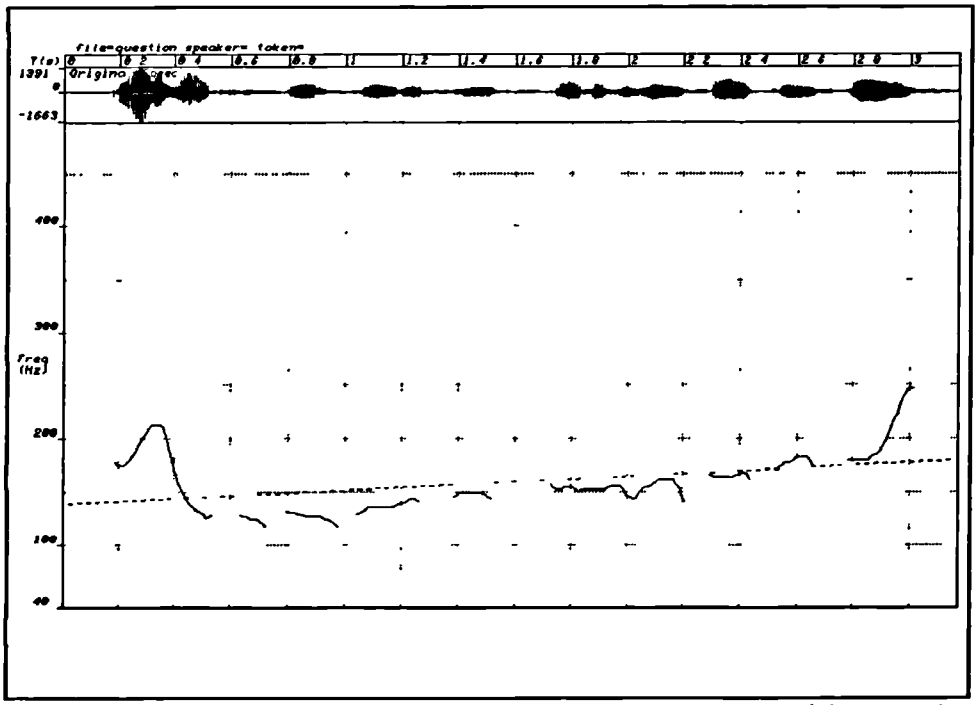


Figure 2.2 Single tone-unit utterance: "ARE rags and bones the only things you can find on the hillside these days?", Speaker M1

it is univariate, not a trivially structured signal.

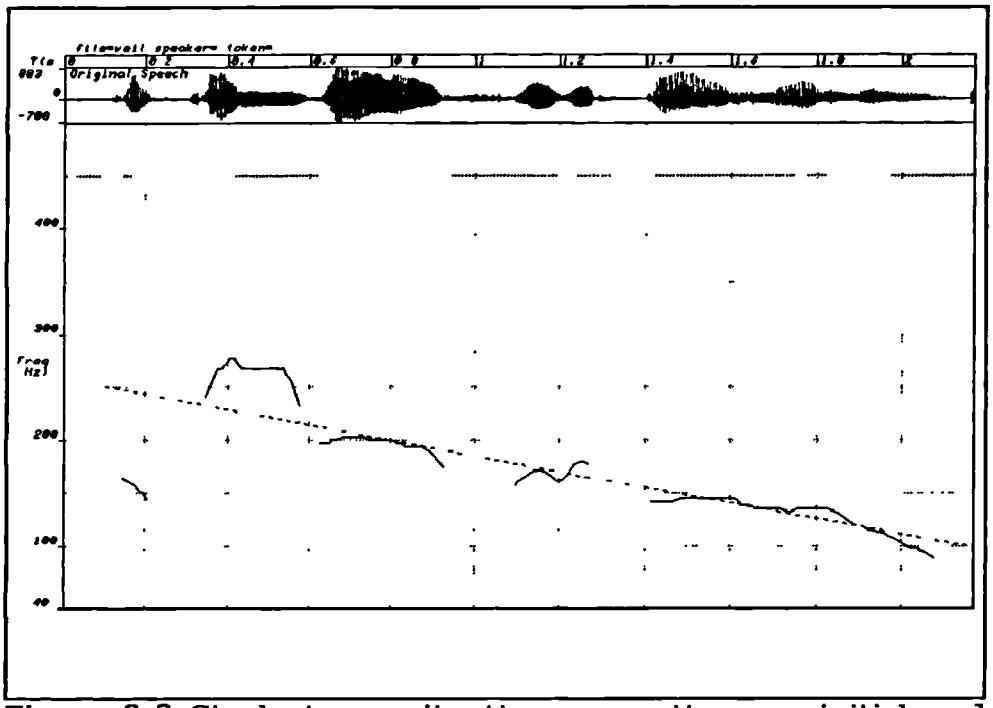


Figure 2.3 Single tone-unit utterance, utterance-initial and utterance-final: "The thin veil slithered down her leg", Speaker M1. Slope of regression line = -73 Hz/sec.

However, there are still some insights to be gained from examining the utility of linear regression. At first sight, fitting a straight line through a F0 contour appears a singularly unpromising method of modelling the elusive phenomenon of declination. It has the advantage of rendering local blips irrelevant, to be sure, but outliers in the contour, such as would appear in utterance-initial accented syllables and utterance-final falling accents, would tend to bias the slope of the declination line in what might be considered to be a misleading way.

For instance, the contour in Fig. 2.3 is taken from a tone-unit which is both utterance-initial and utterance-final. The slope in that contour is clearly much steeper than that in Fig. 2.4a, which is the same clause in utterance-medial position (Fig. 2.4b shows the context from which it was excised). Leaving aside here the question of the domain of declination, (consideration of which might suggest that declination lines should not be computed over any domain less than the utterance or - see discussion of Thorsen's analysis below - that they should be computed over nested domains) it could reasonably be argued that the declination line computed in

Figure 2.3 is conflating the effects of an early boost and a late depression in F0, which are local functions, with those of the phrasing function of declination.

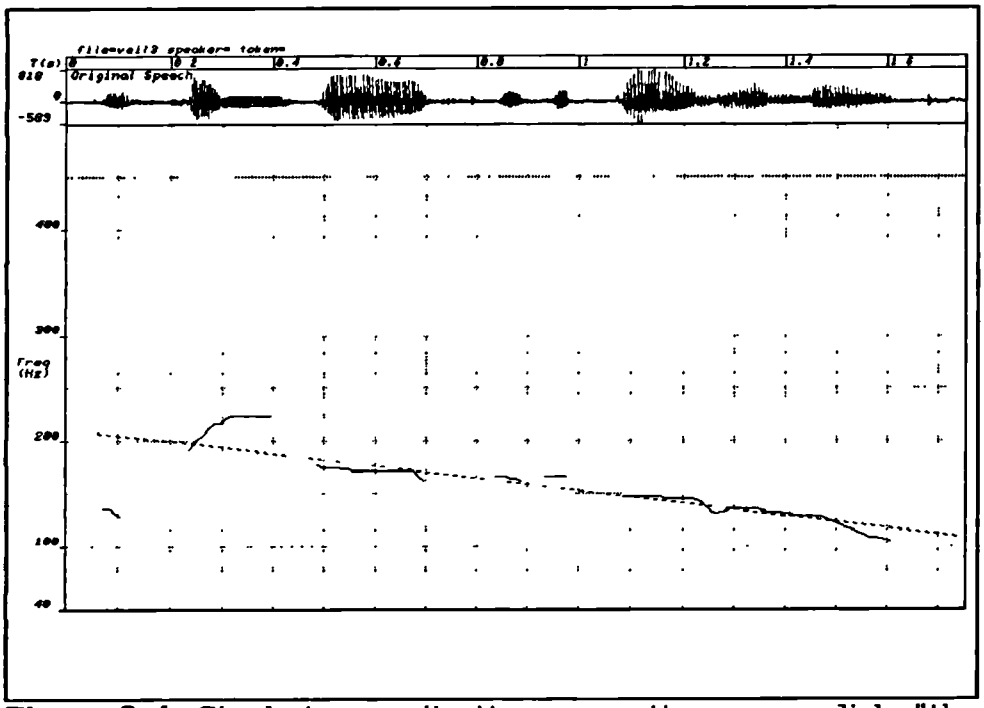


Figure 2.4a Single tone-unit utterance, utterance-medial: "the thin veil slithered down her leg,", Speaker M1. Slope of regression line = -58 Hz./sec.

In fact, the point can be made a more general one against the method of fitting lines by linear regression - that the method takes into account all intonational phenomena in an uninterpreted fashion, such that the effects of other extra-high or extra-low local accents, which could be considered independent of a phrasing function, are incorporated in that function.

A proponent of fitting declination lines by linear regression might object at this point that a declination function should incorporate all the variability present in the F0 contour, because this is precisely what happens to produce the declination effect - variation in degree of accentuation can alter the perceived declination, though not perhaps in a linear fashion. To show that view to be false, it is necessary to show that a 'declination effect' can be present in a contour where the declination line computed by linear regression would not predict it to be present.

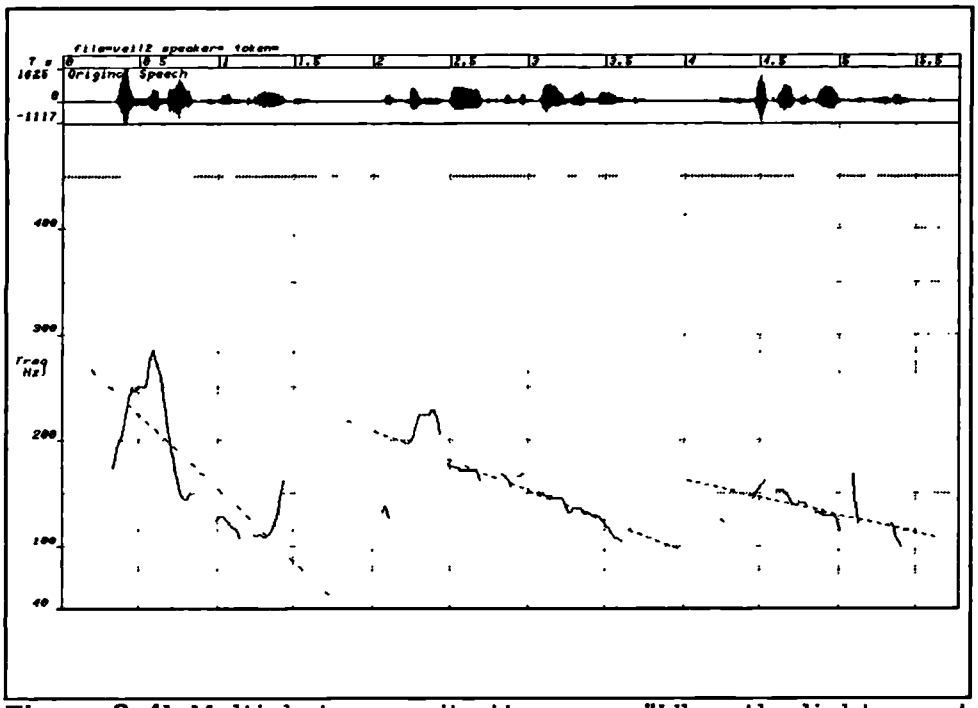


Figure 2.4b Multiple tone-unit utterance: "When the lights went out, the thin veil slithered down her leg, and she let out a gasp of shock.", Speaker M1.

Before addressing that point, it is necessary to be aware of what kind of relationship there might be between a single declination line (computed, for instance, by linear regression) and the 'declination effect'. The declination effect is the phenomenon whereby the pitch prominence of an accented syllable late in an utterance or phrase appears to be equal to that of one earlier in the utterance or phrase when the peak F0 value on the later accent is less by a certain factor than that on the earlier one. For instance, in Figure 2.5, the two accents have peak F0 values which are such as to elicit equal pitch ratings, according to the experimental results of Pierrehumbert (1979) (the difference between the peaks is the 50% crossover point in a labelling curve on a scale of '2nd. peak higher' judgments). The declination effect can be viewed as the expression of a compensatory adjustment in perceived/produced pitch for the declination which is known by speaker-hearers to occur in utterances.

It might be considered reasonable to suppose that the relationship between a single computed declination line and the declination effect is manifested in the following way : when the declination effect is operative in an utterance,

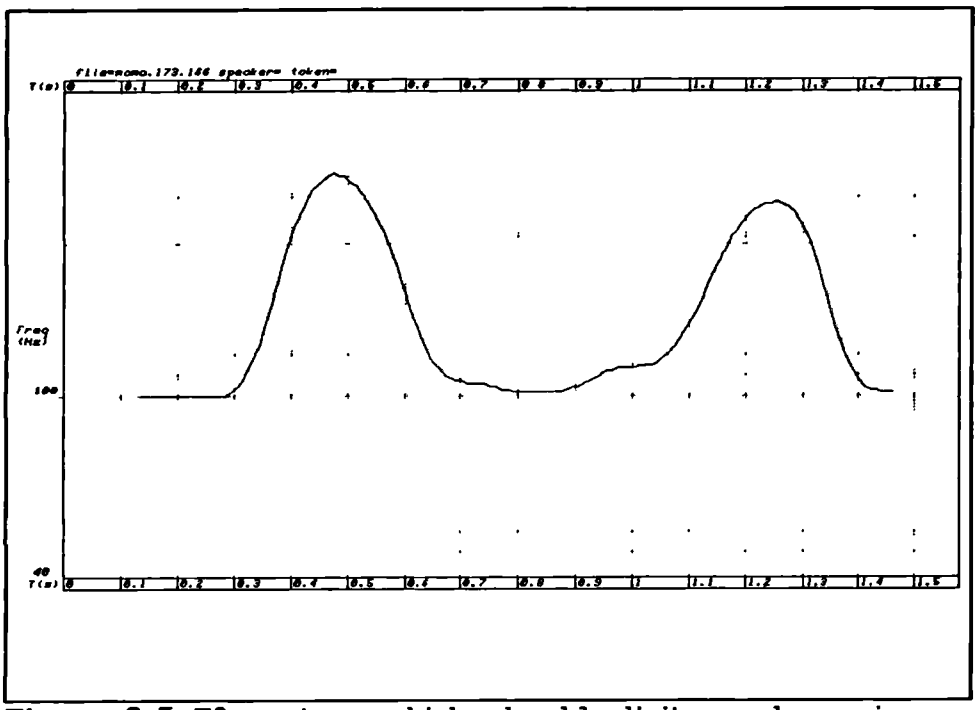


Figure 2.5 F0 contour which should elicit equal prominence ratings for its two accents, according to results in Pierrehumb ert 1979. The frequency difference between the two peaks is 9Hz. The interaccentual duration is 770ms.

there should be a trend line computable from the F0 contour of the utterance which corresponds in some direct quantitative way to the degree of declination effect. The slope of the trend line could indicate the degree to which the speaker-hearer compensates in adjusting the ratio between F0 and prominence in earlier and later accented syllables. However, it can be shown that the relationship between trend line and declination effect can't be as simple as that, even without considering quantitative issues.

The reason this is so is that there are certain F0 contours which have a rising regression line but which elicit the declination effect[6]. The following experiment demonstrates this :

[6] 't Hart (1986) had already suggested the likelihood that a particular type of contour, a long low one with a pointed-hat accent at the end, would have a rising regression line, without investigating the implications.

2.2.2 Experiment 1: testing the validity of the linear regression method

2.2.2.1 Method

The experimental scheme used by Pierrehumbert 1979[7] was adopted, with certain differences. The sentence

The angles were the only things that ever bothered Bill when he was asking Mrs. Jones about the mangles.

(where the underlined syllables are the only two accented syllables of the utterance) was synthesised using segmental synthesis[8], with F0 contours of the form illustrated in Figure 2.6[9]. As in Pierrehumbert's original experiment, two pitch ranges were tested (as a check on the consistency of the declination effect), and there were thirteen different stimuli for the high pitch range (six either side of the stimulus which had the same F0 value on both accented syllable peaks) and eleven different stimuli for the low pitch range (five either side). The F0 values for the peaks in Pierrehumbert's paper were transposed to values more appropriate to the pitch range of the source speaker (i.e. the author) used for PSOLA resynthesis in the earlier

---

[7] In fact, Pierrehumbert's original experiment was repeated in form, if not in number of repetitions, and performed alongside an earlier version of experiment 1, in which reiterant speech was used. Note that reiterant speech was not used in the experiment reported in the main text because of the observations of one of the subjects that it was difficult to imagine an English text which would fit the long rising intonation pattern with a fall at the end. As it was considered that this was likely to be true for most subjects, because that intonation contour is certainly likely to be one that might not be 'stored', being less iconic than many shorter ones, the use of reiterant speech was considered to be likely to affect the results for the long contour, and so a normal sentence was chosen.

[8] The spectral characteristics of the segments were computed according to the Synthesis-by-Rule method of Holmes et al. (1964). The segmental durations were adopted from an annotated natural rendition of the test sentence by the author. 'Steppiness' in the contour resulting from gradual movements through the somewhat quantized pitch levels which are utilised in the synthesis system, was avoided by incorporating a 'moving target' algorithm due to John Holmes, within the quadratic spline fitting routine mentioned in fn. 9.

[9] The crosses mark pitch points, between which quadratic splines have been fitted, as proposed by Hirst (1983). The code for the quadratic spline algorithm was kindly passed on by Daniel Hirst and Robert Espesser at Aix-en-Provence. Hirst and Espesser's algorithm for identifying accentual pivots from a raw F0 contour (Hirst and Espesser 1992) was not used in this case.

experiment mentioned in footnote 7. The transposition was by a factor of 0.5 ERB[10]. The reason that the transposition was done in terms of ERB was Hermes and van Gestel's finding (1991) that prominence relationships are best maintained across registers if an ERB-rate scale is used for scaling the intonation contour, rather than a Hertz or semitone scale.

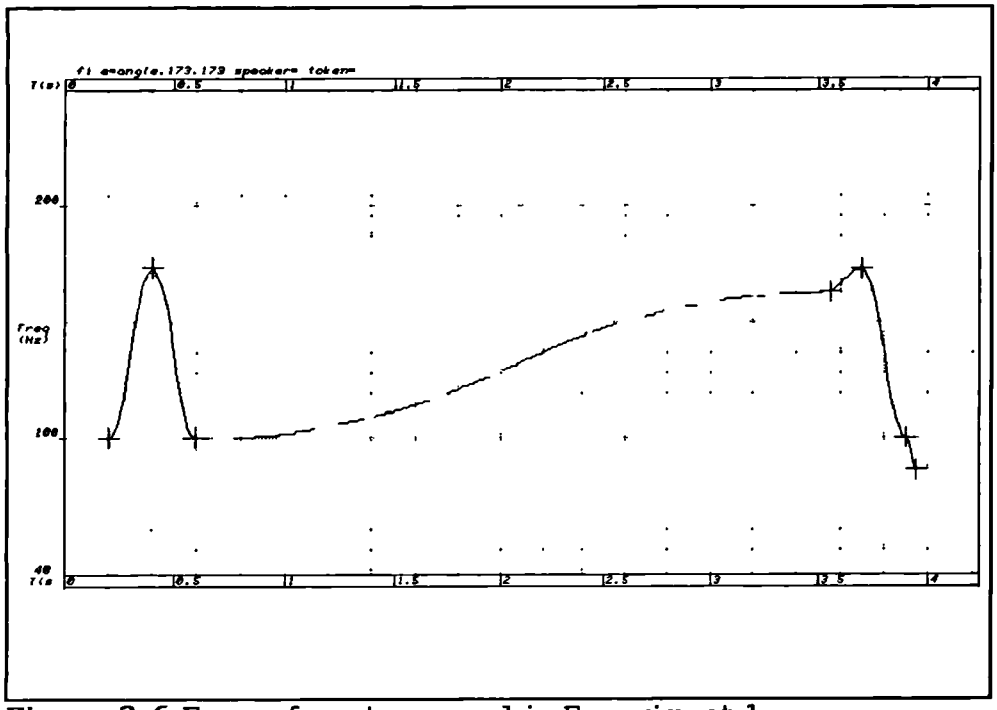


Figure 2.6 Form of contour used in Experiment 1.

Table 2.1 gives the F0 and ERB values of the peaks in the stimuli.

[10] ERB stands for Equivalent Rectangular Bandwidth: A formula for ERB-rate, a scale which Hermes and van Gestel (1991) adopt from Greenwood (1961), is $E = 16.7 * \log_{10}(1+f/165.4)$, where f is the frequency in Hz and E the ERB-rate in ERB.

Table 2.1 Lists of values at which the second peak of the schema in Fig. 2.6 was set in stimuli presented in Experiment 1. Two frequency ranges, wide and narrow; each setting is given in two units (Hz. and ERB).

WIDE FREQUENCY RANGE (1st. peak = 173 Hz, 5.19 ERB)		NARROW FREQUENCY RANGE (1st. peak = 141 Hz, 4.47 ERB)	
2nd. peak (Hz.)	2nd. peak (ERB)	2nd. peak (Hz)	2nd. peak (ERB)
130	4.21	115	3.83
137	4.38	121	3.98
143	4.52	126	4.11
152	4.73	132	4.26
159	4.89	137	4.38
166	5.04	141	4.47
173	5.19	147	4.61
180	5.04	152	4.73
186	5.47	157	4.84
193	5.61	164	5.00
202	5.79	169	5.11
210	5.94		
217	6.08		

Nine subjects were presented the stimuli, which were randomised in two blocks of twelve.[11] They had to rate whether they thought that the first or the second accented syllable was more prominent (the 'accented' syllables were indicated graphically on the question-sheet).

The important thing about the contour family exemplified in Fig. 2.6 is that they have a rising regression line, even when the F0 value on the peak of the second accent is lower than that on the peak of the first (see Fig. 2.7). If

[11] There are no stimulus repetitions in the data, so results are pooled over all nine subject's responses. However, order effects have not been eradicated because the random order was the same for all nine subjects. The lack of repetitions resulted from the intent to perform only a short illustrative experiment in a less than central area of the thesis. However, the question of the effect of the rising interaccentual stretch is pertinent to the central questions of the thesis, and so a fuller version of the experiment, with order effects reduced by the randomisation of stimulus repetitions, is worth repeating in the future. Where order effects have been deemed clearly responsible for anomalous point placement in the plot of Fig. 2.8, they have been discounted from the calculations in fitting the psychometric function, and marked with a cross in the figure.

there were a direct relationship between the slope of the regression line and the declination effect, of the sort 'if non-falling regression line, then no declination effect', then one should expect the declination effect to be nullified or reversed in this experiment.

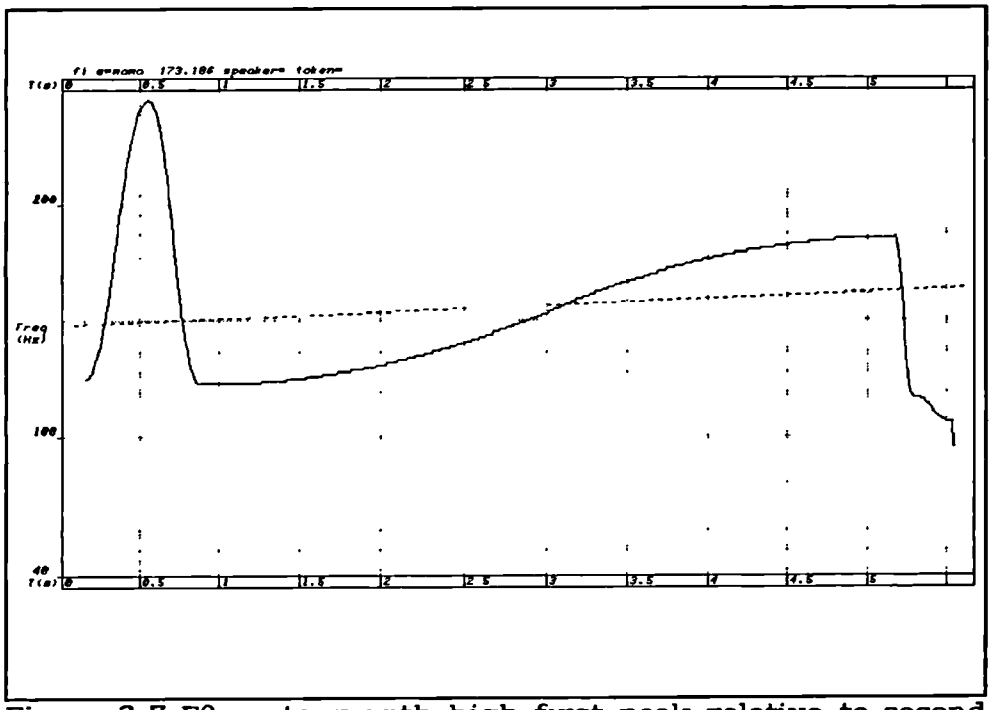


Figure 2.7 F0 contour with high first peak relative to second peak, but with rising regression line.

2.2.2.2 Results

In Fig. 2.8, the proportion of 2nd. peak higher responses is plotted against the difference in peak F0 values between the two accents, values being determined from the pooled responses of all nine subjects. A phi-gamma[12] psychometric curve has been fitted in a similar way as in Pierrehumbert 1979[13]. The Point of Subjective Equality (PSE = the point at which 50% of responses are 'second peak higher') is at the point where the second peak is 9.5 Hz. lower than the first in the high pitch range, and 5.3 Hz. higher than

[12] A phi-gamma psychometric curve is one which conforms to the expectation that categorical responses to stimuli reflect continuous variation, according to a normal distribution, between the different categories.

[13] Note that the distribution-free method of 'jack-knifing' (Mosteller and Tukey, 1977) to discover the true value of a parameter (in this case the 50% crossover point) has not been used in this case. The curve was just fit by Maximum Likelihood Estimation.

the first in the low pitch range. It can be seen for the wide pitch range that the declination effect is operative in this case, despite a rising regression line (the positive crossover point for the narrow pitch range stimuli may have been the result of a response bias, as Pierrehumbert suggests was the case in her experiment).

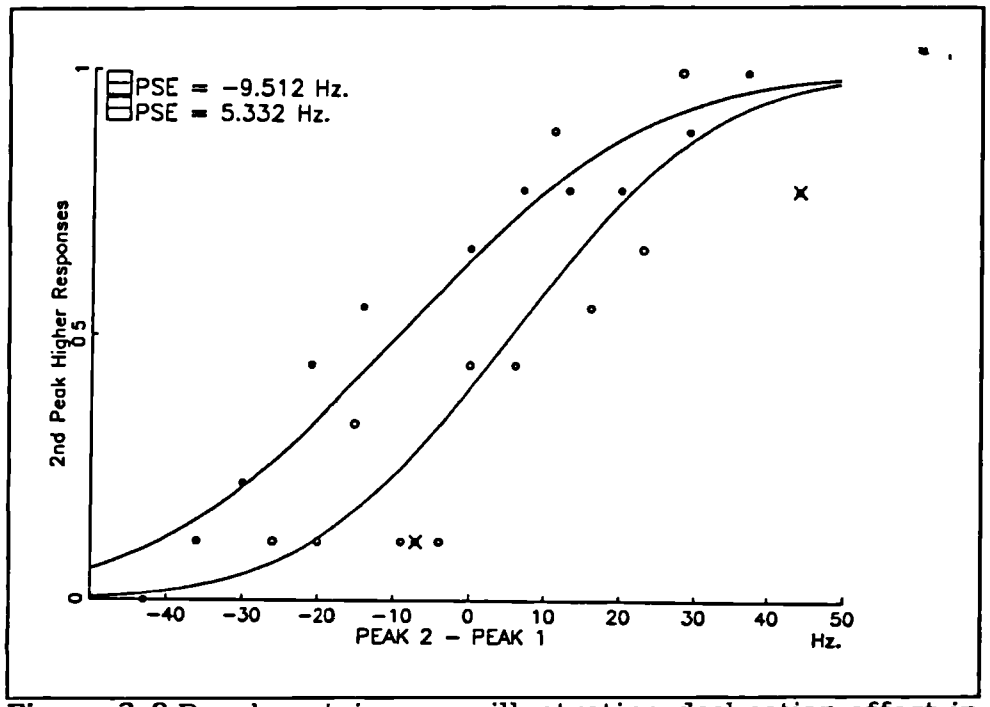


Figure 2.8 Psychometric curve illustrating declination effect in utterances of the form shown in Fig. 2.6.

2.2.2.3 Discussion

This does not mean that a regression line cannot be used to account for the declination effect, but rather that some specific relationship between features of the intonation contour and the regression line would need to be quantitatively expressed for some specific relationship between the regression line and the declination effect to be determined[14]. In the particular case of the contour in Fig. 2.6, a natural candidate for a feature some quantitative parameter of which could be used in conjunction with the

[14] Of course, such a quantitative expression would still have been required even if it had been shown that the declination effect only occurs in the presence of a declining reference line. The experiment merely shows that it cannot just be an additive one between the accent peak F0 values and the regression line values at the position of the accent peaks.

slope of the declination line to produce an estimate of the expected declination effect would be the long rising slope between the accents. For instance, the inter-accentual slope could be hypothesised to have some attenuating effect on the degree of the declination effect. (Given that the pitch range in the high pitch range stimuli is higher in Experiment 1 than the corresponding one in Pierrehumbert 1979, the declination effect of 9.5 Hz in the former, would, on an ERB-rate scale, be less than that of 9.2 Hz. in the latter. Furthermore, if it were true that the declination effect is stronger in longer utterances, the declination effect proportional to interaccentual distance would be less in Experiment 1 than in Pierrehumbert's experiment[15]).

However, it is clear that the regression function would then simply be being used as a subsidiary function in a more complicated model of the relationship between relative accent peak F0 values and the declination effect. In any event, those peak F0 values would themselves have to be used in such a model, so the regression line would be of no use by itself (apart from showing roughly the degree of declination effect expected) unless the distance (in Hz - or whatever frequency unit was adopted) between the point on the regression line marking the middle of the accent peaks was the same as the amount of adjustment in that unit required by the declination effect. But this has just been shown not to be the case for the particular contour-type in Experiment 1. Thus the regression function could only ever be used as a subsidiary function in accounting for the declination effect, and its incorporation might be considered to complicate a declination model unduly, if not just for the reason that peak values would be being used twice within the model, in different ways - once in the regression along with all the other data points, and once in ratio or difference expressions incorporating, for each accent, the accent peak F0 value and the value on the regression line at that position[16].

---

[15] The interaccentual distance in Experiment 1 is 3300ms, and in Pierrehumbert's experiment looks to have been 770ms.

[16] However, it is not out of the question that the features physically present in the F0 contour could have a multiple function, and that that multiplicity is mediated by differing rates and forms of processing that occur at different levels, or in different partitions, of the intonation processing pathways in the central nervous system. Thus accentuation might be

We have just seen that a local[17] phenomenon (a long, interaccentual rising stretch of F0) might have some effect on the prominence scaling of accent peaks. There are some other local phenomena, which could have similar effects. We look at these in turn.

2.2.3 What could be partitioned out?

2.2.3.1 Final Lowering

It is well known that in most, if not all languages, utterances which have a terminal intonational fall descend to a long-term minimum of F0 at the end of them. This minimum is often considered to be a constant for any given speaker (cf. Liberman and Pierrehumbert 1984) and is often referred to as that speaker's baseline[18].

Some examples of final lowering appear in Figures 2.9 - 2.16. In Fig. 2.9, the final syllable of the word 'digits' has a final F0 value (of 115 Hz.) which descends to this female speaker's (speaker=F1) baseline. The lower F0 value is a clear outlier in the distribution of syllable peak F0 values. The same is true of the final syllable of 'letters' in Fig. 2.10 (final value = 80Hz.), for a male speaker (speaker=M2). Often the final syllable of a sentence is uttered with creaky voice (and often with laryngealisation or glottalisation), emphasizing the final-lowering effect. This is the case in Fig. 2.11, on the final syllable of 'seven', although the pitch estimation algorithm (using autocorrelation) has failed to estimate F0 values which correspond roughly

---

physiologically detected using a wide-band filtering system at the same time as declination is monitored using a narrow-band filtering system, using the same batch of auditory nerve impulses as input. In this respect, the discussion of componentiality below is relevant.

[17] By the word 'local' is meant here "whose domain is less than the prosodic domain to which the word 'global' applies". The two terms are thus defined relative to each other. For most of this thesis, the word 'global' will refer to the domain of the tone-unit, and the word 'local' to a domain less than this; typically, to the domain of a single accent, or, as here, of a pre- post- or inter-accentual stretch.

[18] Data in Ladd 1988 and in a normative study by Barry et. al. 19??, however, suggest that this 'constant' speaker characteristic may vary over the long term (a matter of hours or days) within a small range. This should not be surprising, given the vagaries of voice occasioned by fatigue, laryngeal inflammation, nervous tension, and so on.

to the perceived pitch, which is much lower than on the previous minimum, on the word 'score'[19].

In Fig. 2.12, an early nuclear accent (on 'use') allows a more gradual descent to the final low value to be clearly seen in the post-nuclear stretch or 'tail'. Note that there is a subsidiary accent on 'base'. The final rise on the final [n] of 'notation' is a common phenomenon in that position (the author has observed it many times in the course of labelling F0 contours), and appears to be involuntary. Fig. 2.13 exemplifies another type of rise that can occur after final lowering (on the syllable 'rare', here spoken by a female, F2). Again, it appears to be involuntary, is not large enough to have any phonological significance, and is probably the result of target overshoot (the low pitch reached could not be maintained without descent into creak, which is not appropriate for a non-sentence final fall; note too that the terminal rise in F0 that results is not perceived as being phonologically significant (i.e. it is not what is referred to in the literature as a continuation rise) because of the low speech amplitude with which it is uttered). In Fig. 2.14, creak is adopted, because it is highly appropriate in the context, which is evident from the text. Fig. 2.15 demonstrates a particularly good example of the way in which final lowering can mark an abrupt termination of a gradually declining trend (female speaker F2). The same speaker demonstrates the generality and consistency of the phenomenon in Fig. 2.16 by terminating an utterance of the same text in the same way as speaker F1 in Fig. 2.9.

[19] Evidence for creak and laryngealisation in this and other passages is found in the Lx waveform which is coterminous with the speech waveform in the data of all these passages, taken from the Eurom-0 database prepared during the Esprit SAM project.

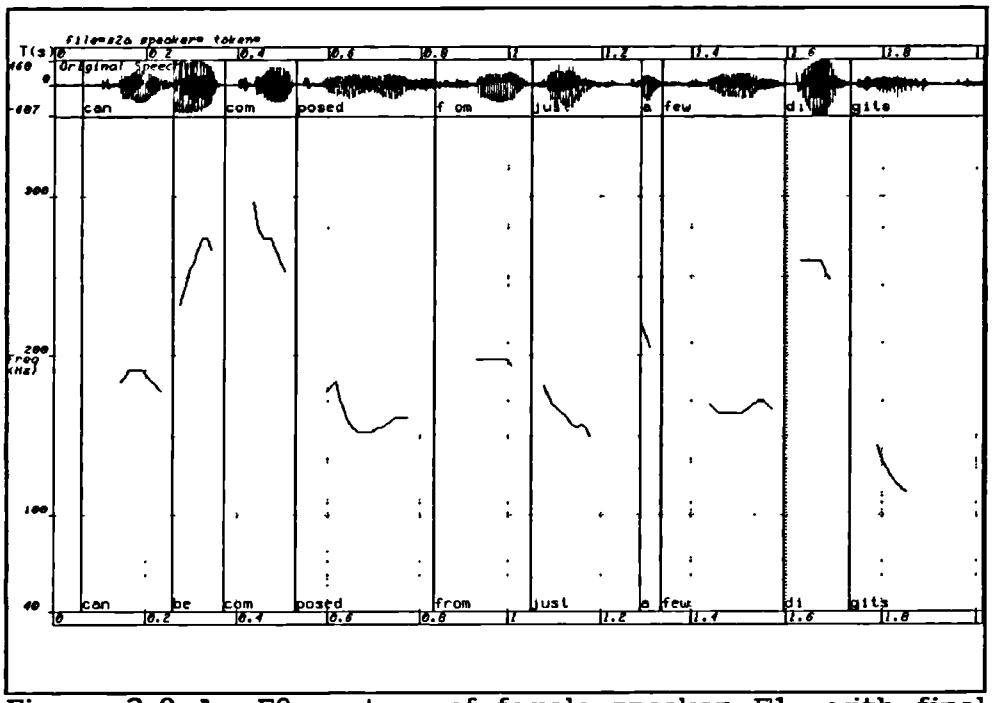


Figure 2.9 An F0 contour of female speaker F1, with final lowering of the contour on the second syllable of "digits'.

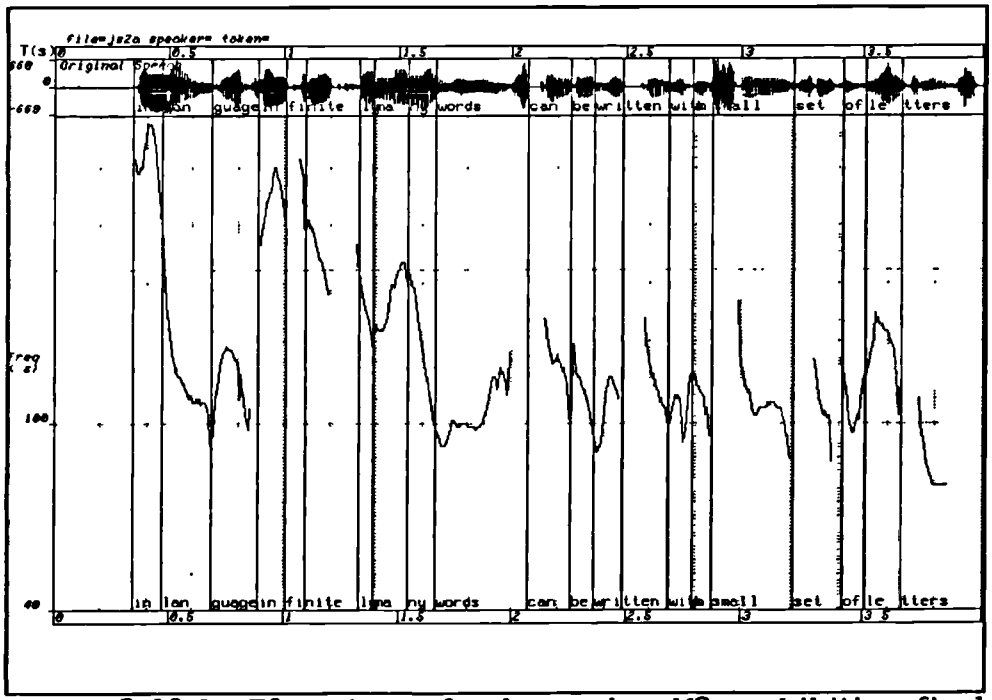


Figure 2.10 An F0 contour of male speaker M2, exhibiting final lowering on the second syllable of 'letters'.

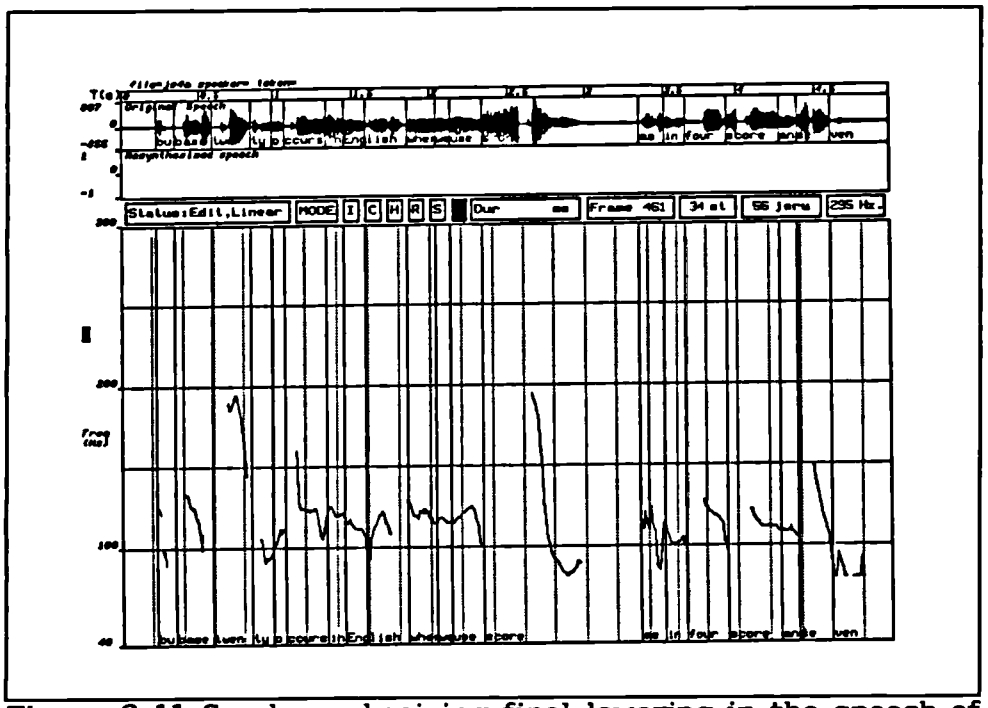


Figure 2.11 Creak emphasizing final lowering in the speech of male speaker M2 (on the second syllable of 'seven').

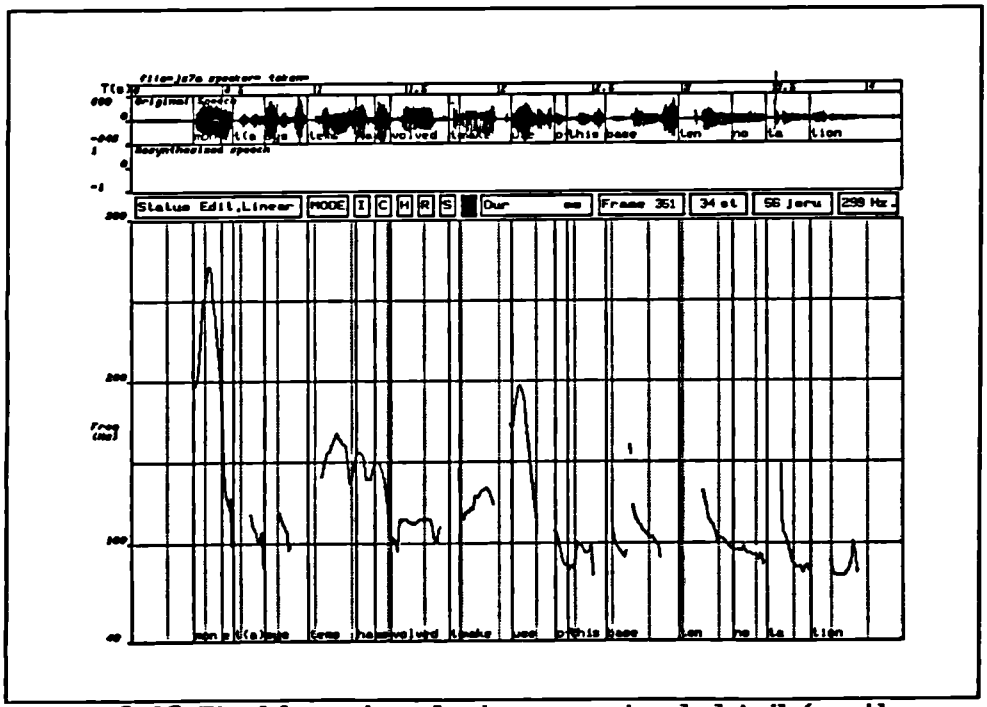


Figure 2.12 Final lowering during an extended tail (on the words 'ten notation'). Speaker M2.

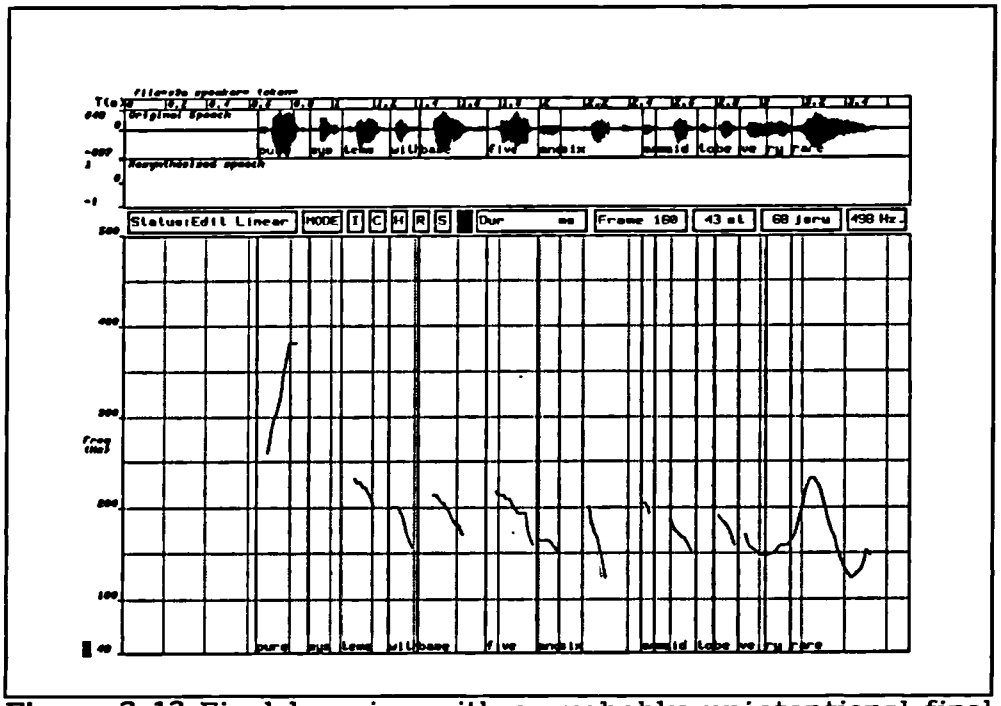


Figure 2.13 Final lowering with a probably unintentional final rise on the word 'rare'. See text for explanation. Speaker F1.

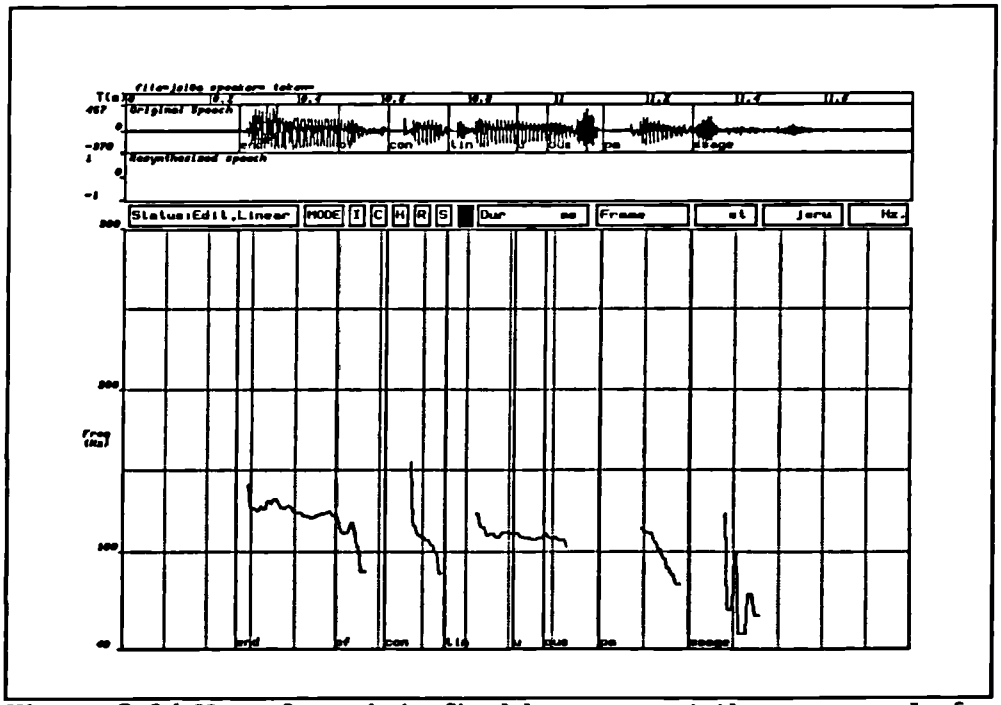


Figure 2.14 Use of creak in final lowering at the very end of a spoken passage (on the second syllable of 'passage'). Speaker F2.

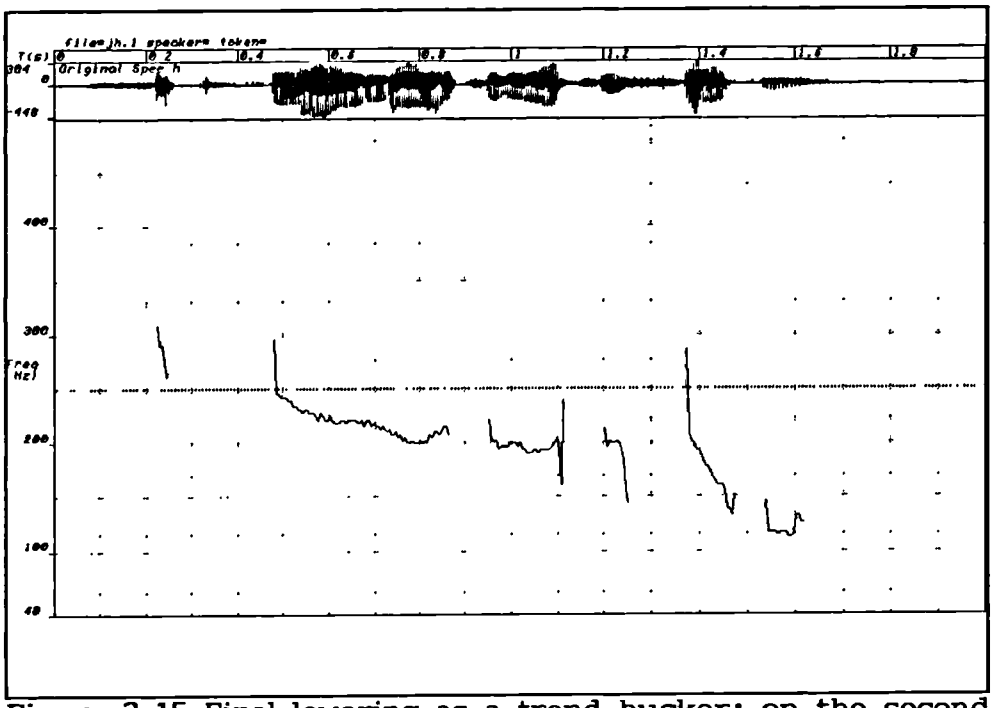


Figure 2.15 Final lowering as a trend-bucker; on the second syllable of 'seven'. Speaker F2.

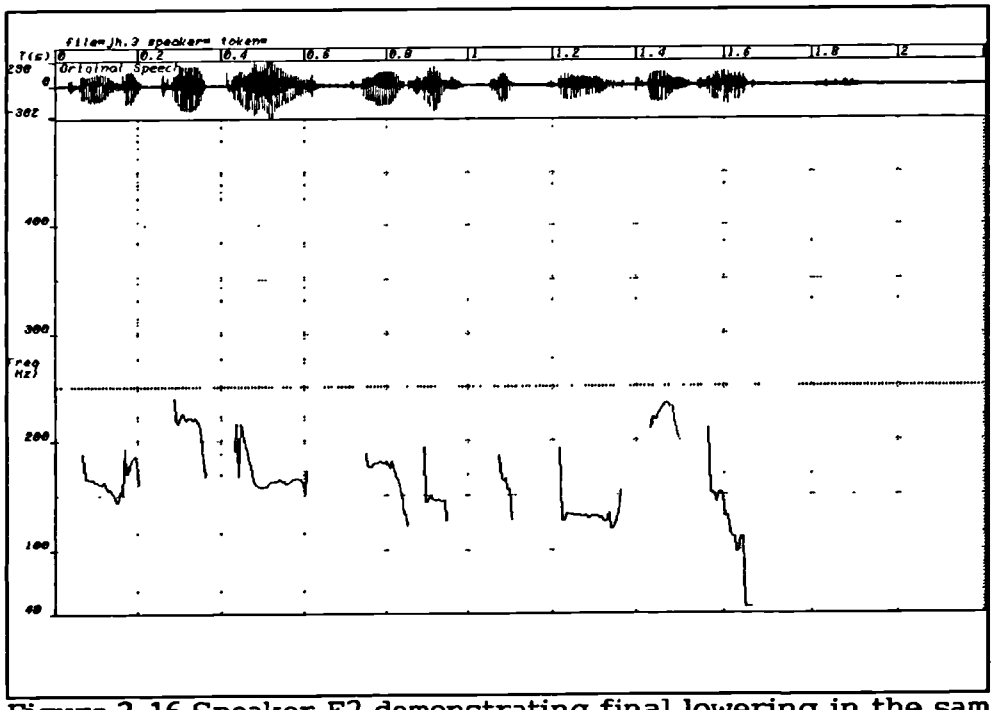


Figure 2.16 Speaker F2 demonstrating final lowering in the same way as Speaker F1 (see Fig. 2.9).

The major points of issue concerning Final Lowering do not include whether it exists, but are (i) how long it lasts, and (ii) how the F0 contour reaches the final value : suddenly or gradually. In respect of (i), Lieberman et al. (op.cit.) give a rather brief estimate of 150 ms. for the phenomenon. Liberman and Pierrehumbert (1984) consider it to apply to the last accent in an utterance, but given that the last accent in the utterances they have analysed is always fairly close to the end of that utterance, it is difficult to judge from their data whether the phenomenon is something local to a phonological 'event' (the final nuclear accent) or something which occurs over a short space of time at the end of an utterance, regardless of the placement of accents. Silverman (1987) makes this point and produces evidence which suggests that the effect of final lowering can occur on a number of syllables close to the end of an utterance, with diminishing effect the further from the end of the utterance they are. This treatment of final lowering is also adopted by Pierrehumbert and Beckman (1988) (see below) who also observe that the effect of final lowering observed in Liberman and Pierrehumbert (1984) was of the order of half a second. Gussenhoven and Rietveld (1988) refer to the same amount.

Physiological evidence suggests that the phenomenon can be viewed in terms of a terminal adjustment that is made at a variable point at or after the last accented syllable, and that its onset results from the interaction of the effects of being in a particular position in the pitch range and being at a particular temporal point in the intonation contour. In Fig. 2.17, taken from Collier 1974, are some physiological traces associated with particular Dutch intonation contours. It can be seen that in all these declarative utterances, there is a nexus of physiological activity, comprising simultaneous cessation of cricothyroid muscle activity (which is associated with pitch-raising), rapid onset of sternohyoid activity (which is associated with pitch-lowering), a sudden increase in the rate of subglottal pressure decline, and a rate of vocal fold vibration corresponding to a particular fundamental frequency value (here about 120Hz). If there is sentential material available to continue the utterance, there then follows a stretch of sustained lower-level but supra-baseline sternohyoid activity, clearly associated with maintaining a low pitch, along with continued decrease in subglottal pressure at the new rate of decline until the end of phonation. Interestingly, the only exception to this state of affairs appears to be contour 14, in which the two muscular

conditions coincide with a downstepped accent in the intonation contour (at about 120Hz.), and in this case there is no sudden increase in the rate of subglottal pressure decline. It is coincidentally interesting also that the value of 120Hz. noted in this analysis corresponds to the upper bound of Atkinson's (1978) Mid F0 range for a male speaker, within which there is highest absolute correlation between variation of a physiological variable (SternoHyoid activity) and F0 variation.

Of course, this doesn't enable us to establish any rules for the location and shape of final lowering when looking at the F0 contour alone. It would seem that on this evidence, the best we could do would be to determine the pitch range of an individual, decide by analogy with physiological studies such as Collier's where the point in the pitch range is that the said nexus of activity is likely to occur, and hypothesize that final lowering occurs once that point in the pitch range is passed and all subsequent F0 variation is downward. At the same time, it should be borne in mind that Collier's study is performed on Dutch utterances, and there is a suggestion (Terken 1989a, Kraayefeld, personal discussion) that there is no such thing as Final Lowering in Dutch. This suggestion is corroborated by the fact that, as we shall see, Dutch models of Dutch intonation have no difficulty in accounting for such localised downward variation in F0 within the framework of declination for which, inter alia, they have become famous.

Evidence from another language, on the other hand, suggests that the phenomenon of final lowering can be even more pervasive than suggested by the physiological evidence. Pierrehumbert and Beckman's (1988) study of Japanese intonation includes a comparative analysis of question and statement contours on the same text which suggests that final lowering might start earlier in an utterance than has just been suggested; specifically, some time before the final accent. Fig. 2.18 is a duplicate of Pierrehumbert and Beckman's Fig.3.9, in which the comparison between mean salient F0 points between question and statement contours is displayed for two speakers. For the first speaker, there is a divergence between the two which suggests the onset of final lowering in the statement four syllables before the final pitch accent, with no final lowering in the question. For the second speaker, the lowering effect appears to start at the very first syllable, for which effect the term 'phrasal lowering' would probably be more appropriate. The

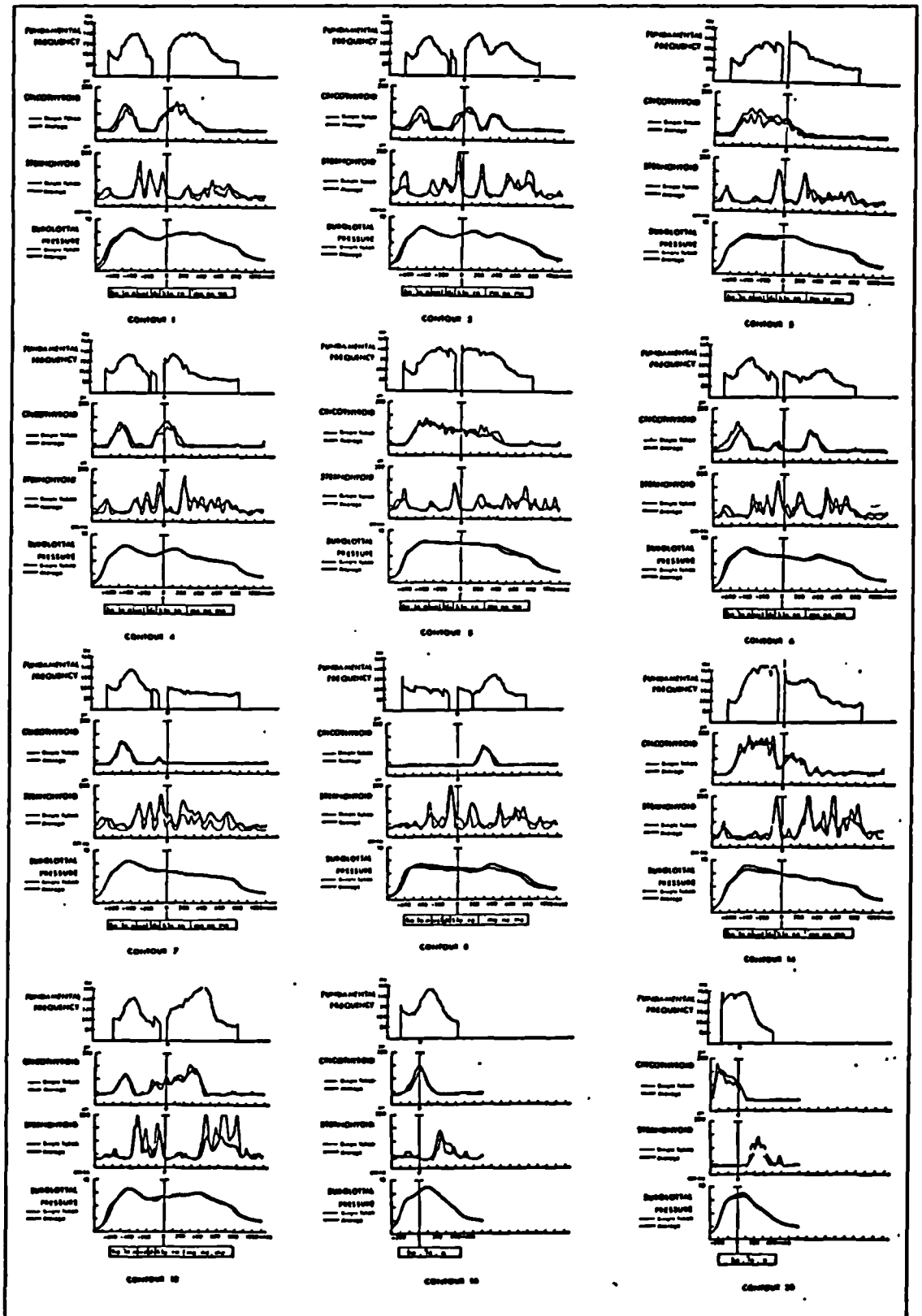


Figure 2.17 Physiological traces taken from the Appendix to Collier 1974. All but his contour 14 demonstrate a particular pattern of physiological activity at the onset of Final Lowering.

question naturally arises to what extent the divergence in the contours shown is a function of raising caused by the need to communicate the pragmatic function of questioning. Notwithstanding that issue, the effect one way or the other (and it is probably a bit of both) is clear in those illustrations, and perhaps representative of a significant difference between the two contour types. If that is the case, then (at least for Japanese) a certain degree of 'final' lowering can occur at a more variable position than had previously been suggested, even to the extent that it becomes not a local, but a global lowering phenomenon[20]. However, judged from those illustrations, the rate of lowering still increases to a maximum at the end of the utterance.

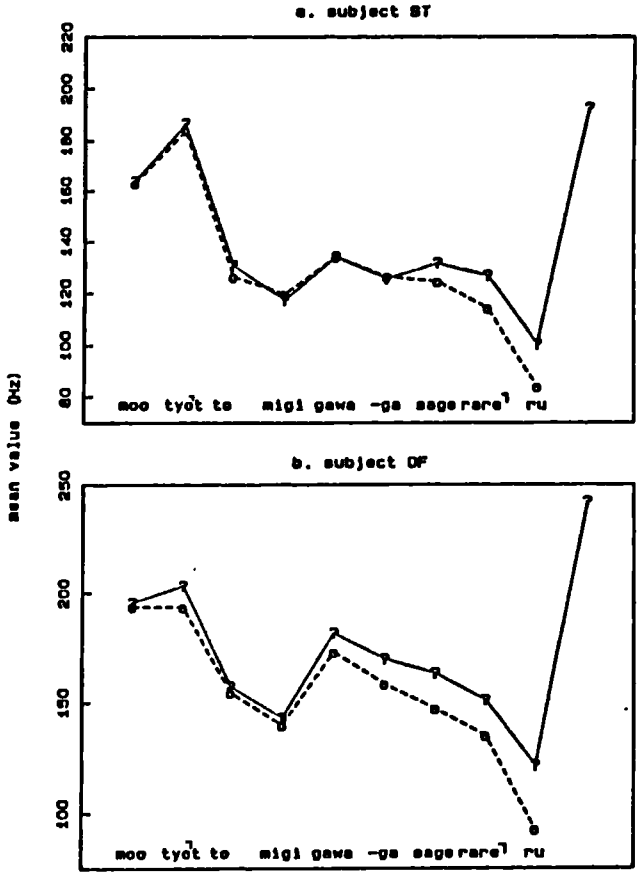


Figure 2.18 Taken from Pierrehumbert and Beckman 1988, Fig. 3.9, p.74. Mean values at each measurement point in utterances of "Moo tyo'tto migigawa-ga sagerare'ru" produced by (a) their speaker ST and (b) by their speaker OF.(o=statement,?=question).

The shape of the lowering function, in all the illustrations in this subsection, appears to be a linear downward slope, rather than a sudden switch to a

[20] However, it is treated as a phenomenon restricted to declarative utterances.

lower level. In this respect, it is more akin to local declining stretches of F0 than, say, downstep (see below). For this reason, it might be supposed that there is nothing categorically different from declination in final lowering, and that the phenomenon is better modelled as a (perhaps non-linear) adjustment of the declination line, whatever form that takes. This might be a natural thing to do if the F0 contour is considered alone (and, as indicated, is what the Eindhoven school does in modelling the intonation of a number of languages). However, the physiological evidence adduced above suggests that a more categorical switch in conditions likely to affect the mode and rate of vocal-fold vibration takes place at a particular point in a speaker's pitch range (and perhaps only in respect of certain intonation patterns), and for that reason the suggestion is still valid that final lowering is a specifically local phenomenon. As we shall see, that does not preclude it being modelled as part of a global trend.

2.2.3.2 Initial Raising

Just as a local phenomenon can be seen to buck a trend at the end of an utterance, so there is one which appears to do the same at the beginning of an utterance. This is the phenomenon of Initial Raising[21], which occurs most markedly in that position, but could also arguably be considered to occur utterance-internally, at the beginning of utterance-medial and utterance-final tone-units. Figs. 2.19-21 demonstrate the phenomenon. In Fig. 2.19, the height of the initial

[21]This term is used in a number of ways in the literature. It has been connected with an upstep mechanism which only operates under certain metrical conditions (Kubozono, 1989, Ladd and Johnson, 1986). It was also used by Silverman (1987) to describe an upward range reset at the beginning of intonational paragraphs. Here it is just a phenomenological gloss.

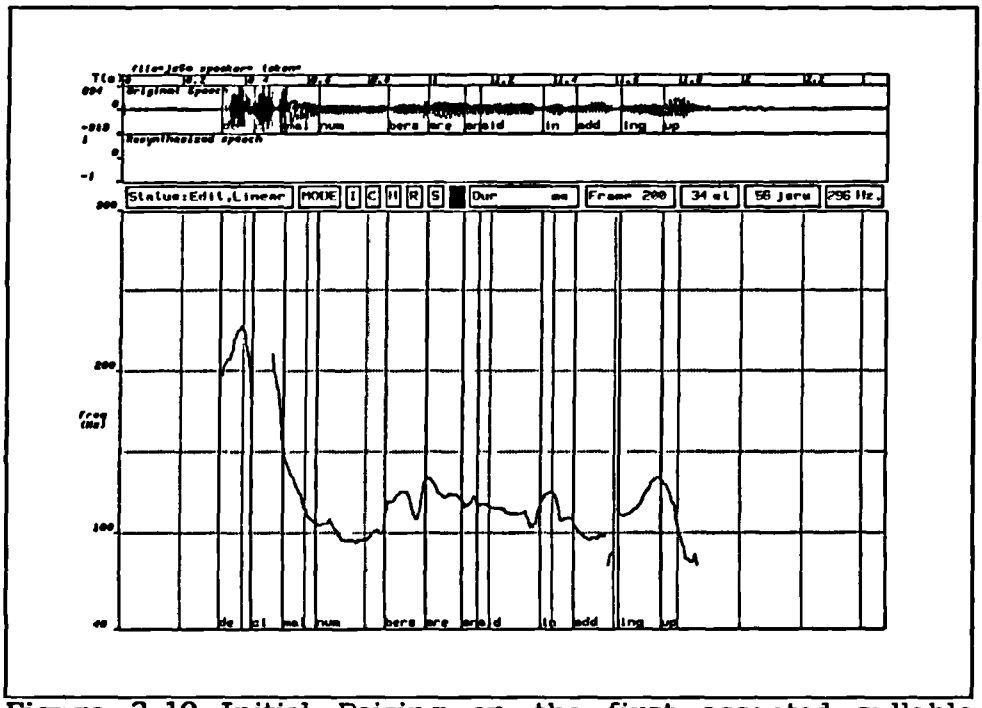


Figure 2.19 Initial Raising on the first accented syllable ("decimal") of a paragraph; Speaker M2.

accent is clearly greater than that of those in the rest of the utterance. The utterance is paragraph-initial. The increased F0 value on the accented syllable peak is, however, likely to be not just a result of its textual position, but also of the fact that the word "decimal" is being contrasted against descriptions of other number systems. In Fig. 2.20, there is no such conflation of causes, but the value of F0 on the initial accent (on the word "try") is of the same order as that on the initial accent in Fig. 2.19. In Fig. 2.21, the initial accent (on the third syllable of "Scandinavian") is not paragraph-initial, but is significantly higher than that on the first syllable of "Russia"; it should be noted that the latter is not in a position to undergo final lowering. This is a possible example of utterance-internal initial raising.

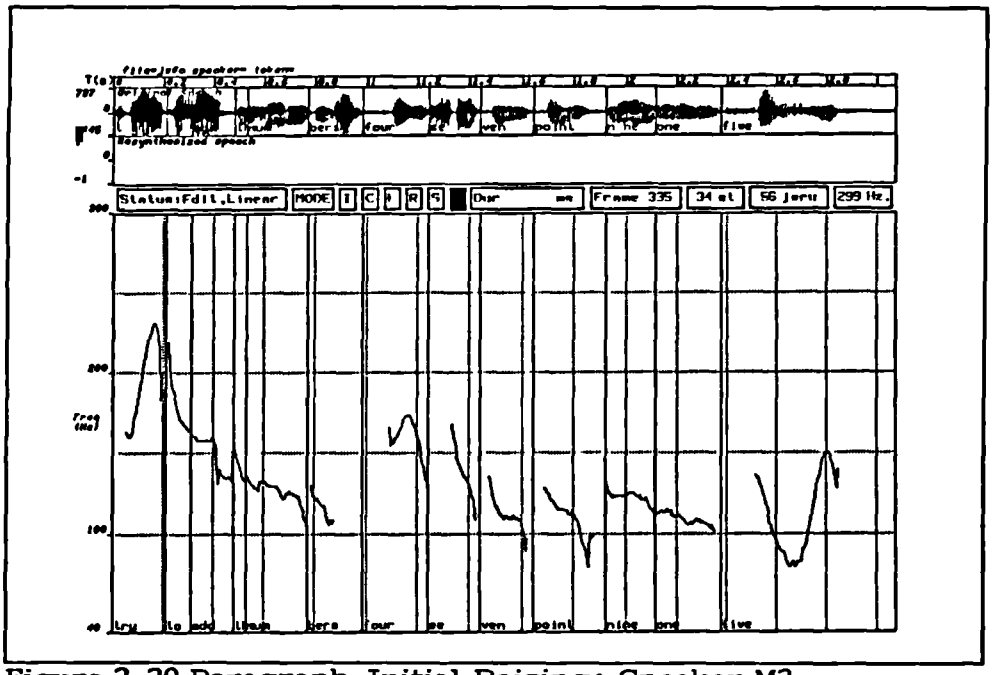


Figure 2.20 Paragraph-Initial Raising; Speaker M2.

The phenomenon of initial raising is typically considered to apply only to the first accented syllable of a prosodic unit. Thus, it is not considered to be a function necessarily similar to that of final lowering, which, as has been suggested, tends to occur regardless of the accentual structure of an utterance. There are, moreover, at least two facets to the phenomenon. Firstly, an accented syllable is likely to be uttered with increased F0 when it is at the beginning of a prosodic unit (be it intonation phrase, breath-group or paragraph), such that the absolute difference in F0 between that first accent in the first prosodic unit at the next level down in the prosodic hierarchy and the first accent in the second such prosodic unit is greater than that between the first accent in the second such prosodic unit and the first in a third such prosodic unit[22] [23]. Secondly, an

[22] On this interpretation, Initial Raising would be deemed to occur within the domain of the breath group if the difference between the F0 on the first accent in the first Intonation Phrase (IP) of the breath-group (BG) and that on the first accent in the second IP in the BG was greater than that between the F0 on the first accent in the second IP in the BG and that on the third IP in the BG. Similarly, Initial Raising would be deemed to occur within the domain of the IP if the difference between the F0 on the first accent in the first accentual phrase (AP) of the IP (i.e. the first accent in the IP, because there is only one accent per AP) and that on the first accent in the second AP of the IP (i.e. the second accent in the IP) was greater than that between the F0 on the first accent in the second AP of the IP (i.e. the second accent in the IP) and that on the first accent in the third AP of the IP (i.e. the

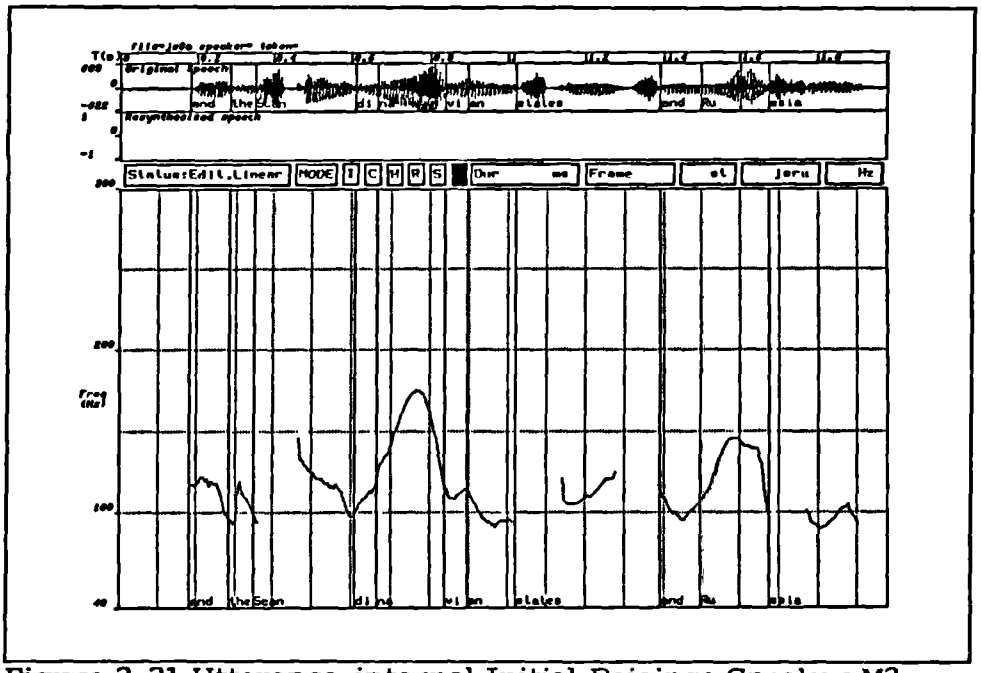


Figure 2.21 Utterance-internal Initial Raising; Speaker M2.

accented syllable at the beginning of a prosodic unit can be uttered with increased F0 if the duration (measured by the number of syllables in the unit) increases.

There is some debate over whether this second facet is consistent in its manifestation. Cooper and Sorensen (1981) found that the difference in F0 between the initial accent in an utterance and its successor was markedly

---

third accent in the IP). This latter case can be more succinctly put as follows:

In the domain IP,

Initial Raising is taken to occur if
$A_1-A_2 > A_2-A_3$
(where A_i is the peak F0 value on the ith accent).
This inequality must be considered independently of any of the effects of downstep (see section 2.2.3.3).

In the absence of a third element within the relevant domain, and of any additional contextual clues, Initial Raising cannot be taken categorically to have occurred, since an apparently large difference between the peak F0 on the first and second relevant accented syllables might then be attributable (ceteris paribus) to an increase in pitch range over the relevant domain.

[23] For a treatment in which the domain of Initial Raising is restricted to the paragraph, see Silverman 1987.

higher than that between the second and its successor, such that the model they decided on in accounting for their sentence intonation data treated the height of the initial accent as a special phenomenon, and predicted only the value of F0 after the first accent. They also found (in their experiment 2.1.1) that an increase of the length of an utterance (by a factor of 2) corresponded with an increase in the value of F0 on its initial accent (by a factor of 1.06). They noted (p.38) that a similar phenomenon had been noted by McAllister (1971) and O'Shaughnessy (1976), but that an experiment using list utterances rather than renditions of sentences of varying syntactic complexity (Cole at al. 1980) had not produced such a length-dependent increase in the F0 on the initial accent. Similar results for list utterances were found by Liberman and Pierrehumbert (1984).

Ladd and Johnson (1986) have pointed out that the results from a number of studies are inconsistent on this point, and suggest that that aspect of initial raising which appears to depend on utterance length is in fact triggered by a branching metrical structure in the constituent bearing the initial accent, in particular such that that constituent is right-branching (the main accent being on the right) so that the preceding accent has to be uttered with increased F0 for the targetted F0 on the second accent to be attained. The experiment they conduct to test this hypothesis only bears it out for one of two speakers tested, however.

2.2.3.3 Downstep

The first two of the local phenomena discussed in sections 2.2.3.1 and 2.2.3.2, final lowering and initial raising, occur at the periphery of a prosodic unit or utterance. The third occurs in a medial position within such domains, and at the same time is, in the opinion of this author, of somewhat less clear status. The phenomenon of downstep was incorporated into a phonological account of the intonation of English by Pierrehumbert in her thesis of 1980, and used, inter alia, to account for those English intonation patterns in which the peak F0 of successive accented syllables decreases asymptotically. Some of these contours (the ones containing a 'stepping head' in O'Connor and Arnold's (1973) and Kingdon's (1958) parlance) have been claimed (by those same authors) to be the most common in English utterances, though it would seem to the current author that less precise styles of modern speech have increased the predominance of 'falling heads'.

Others are less common (e.g. Crystal's 'spiky head' (Crystal 1969), O'Connor and Arnold's 'sliding head'(O'Connor and Arnold 1973)).

In Figs. 2.22 and 2.23 are two examples of downstep in a stepping head. The contour in Fig.2.22 has already been used in Fig. 2.20 to demonstrate Initial Raising. In the utterance 'Try to add the numbers four seven point nine one five...', the accent on the syllable 'add' has been downstepped relative to that on the syllable 'try', and that on 'num' downstepped relative to that on 'add'. This shows that the process is iterative. In Fig. 2.23, in 'only ninety six years after that...' the accent on the syllable 'six' has been downstepped relative to that on 'nine', and that on the syllable 'years' downstepped relative to that on 'six'. It is of interest and importance to note that a fourth accent, on 'af', has not been downstepped relative to its predecessor. This demonstrates that downstep is not a mandatory process on all successors to an accent in a prosodic domain which has undergone it - at least in the treatment which instigated its incorporation into a phonological account of English intonation, an analysis of which appears in chapter 3.

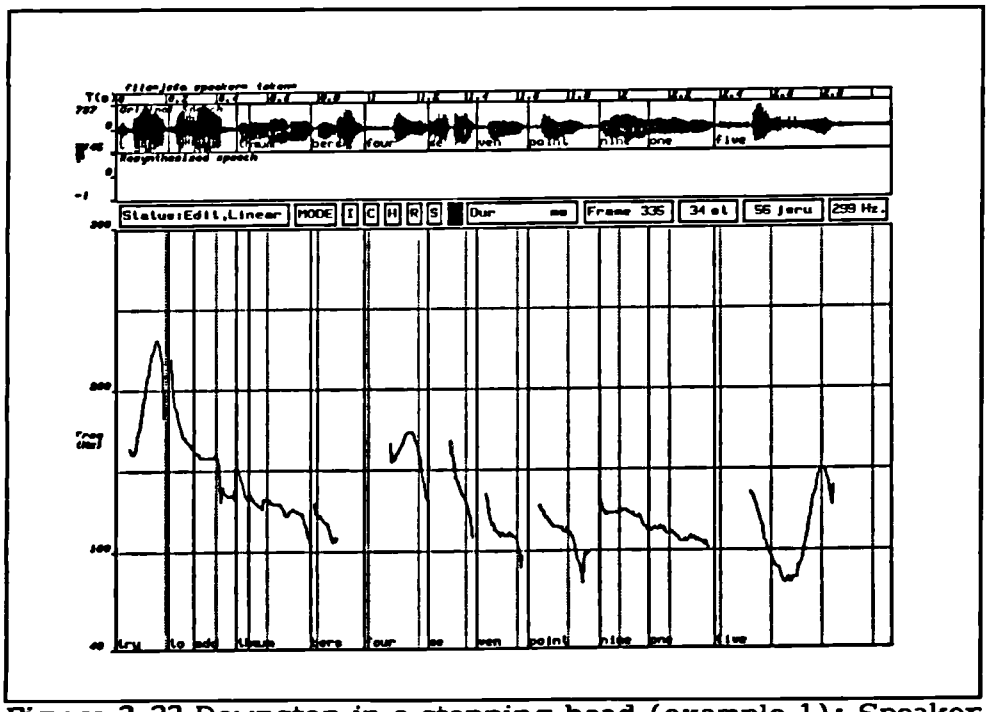


Figure 2.22 Downstep in a stepping head (example 1); Speaker M2.

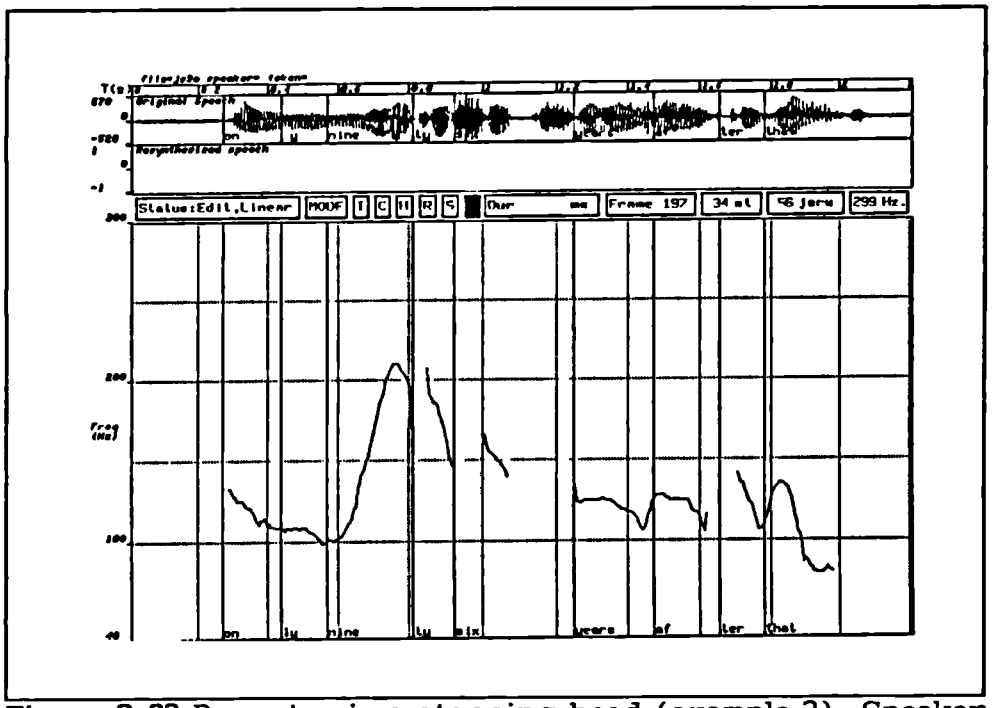


Figure 2.23 Downstep in a stepping head (example 2). Speaker M2.

2.2.3.4 Conclusion

This brief exposition of local phenomena, which is extended in chapter 3 in respect of the phenomenon of downstep, has examined the question of what of a downtrend, in at least an English intonation contour, can be attributed to local variations, and what to some global, slowly declining component. The issue is not yet fully resolved. Final Lowering is a local phenomenon that could be taken to occur in most languages, as it would seem to result from relaxation of the phonatory and respiratory musculature at the end of an utterance. Initial Raising might be equally universal, resulting from the often high emphasis that an initial accent is spoken with at the start of a prosodic domain when it is not scaled relative to predecessors. However, Ladd has suggested that the existence of Initial Raising depends on metrical structure, and there is some evidence that this could be the case.

As for downstep, the brief introduction to the phenomenon in this chapter has merely shown that it is possible to identify some contexts in intonation contours in which it occurs[24]. Here we withhold judgment on downstep sequences other than the canonical form, a stepping head. The possibilities

[24]to wit, within iterative patterns

of local isolated instances of downstep occurring, and the canonical form being reinterpreted, are explored respectively in Chapters 3 and 5.

It is not even clear that once one has discounted all the local phenomena from analysis of downtrends in intonation contours, that one is left with a global slowly-declining trend. There is still the possibility that what remains is another local function, this time a local declination function operative only in unaccented stretches of speech (this hypothesis is explored in Chapter 4). That possibility, if only because no purely local features are taken by them to be involved in downward trends in intonation, would certainly be dismissed by the group of analysts to be discussed next.

2.2.4 Accounts of declination as a frame of reference

2.2.4.1 The Eindhoven School

There is a way of accounting for those aspects of the downward trend in intonation contours which pertain to accented syllables (or the last fraction of a second of an utterance) which is different from attributing them to local functions. This is to account for all such phenomena by generating intonation contours within a moderately steeply declining framework. This is the approach that has been taken by researchers at IPO, Eindhoven since the early to mid-sixties, the tenets of whose work have generally gone under the rubric of the 'Dutch School'. This work, and the researchers engaged in it, is referred to here as the 'Eindhoven School' to distinguish it from work elsewhere in the Netherlands which adopts some of the analytic methods from IPO but conducts independent lines of research (such as at the Catholic University of Nijmegen).

Generating intonation contours within a declining framework involves devising a formula for either (i) a declining baseline relative to which all pitch movements in the contour are computed (ii) a declining topline relative to which all pitch movements in the contour are computed, or (iii) both (i) and (ii) together. It is a well established fact that pitch movements in an intonation contour have more scope for excursion at the top of the pitch range that a speaker has temporarily adopted for his/her utterance (assuming they are not speaking at the top of their available range) than at the bottom of that range. This is because speakers generally speak only towards the bottom of their available range (Crystal 1969, p.111).

Consequently, it makes more sense to use a declining baseline as a reference point relative to which pitch values are computed than a declining topline[25]. The Eindhoven school, however, use option (iii) of those just presented; that is, intonation contours are computed relative to a declining baseline and topline. Not only that, but, with a few exceptions, the pitch movements that comprise the intonation contours all move between or follow the line of these declining reference lines. That is, the declining topline and baseline not only act as abstract reference lines for the pitch movements but also as physical reference points and as part of the intonation contours themselves. How this is, and the reason they can use both topline and baseline as the declining framework, will be seen more clearly in the following sections.

2.2.4.1.1 The Eindhoven school analysis of Dutch

The Eindhoven school has adopted a specific strategy in accounting for the large amount of variation seen in intonation contours due to degrees of accentuation and segmental coarticulation effects. This is to establish, by a process of analysis-by-synthesis, the perceptually relevant pitch movements in an intonation contour, to draw up a minimal family of such pitch movements, and then to analyse all such intonation contours in terms of such movements. All and only the intonation contours of Dutch can then be generated by a grammar in the form of a transition network generating sequences of such pitch movements. By this procedure, they aim to solve the problem that to develop phonological models of intonation on the basis of auditory evidence (on the part of the researcher) allows the possibility of important phonetic variation being omitted, whilst to develop phonetic models of intonation on the basis of instrumental evidence allows the possibility of redundant variation in the speech signal overwhelming the variation that is significant for a listener ('t Hart et al. 1990, pp. 2-4). This approach suggests a perceptually-oriented interpretation of communicative processes. In fact, it incorporates other assumptions which suggest that it is an approach consistent with the Motor Theory of speech perception (see p. 76), although it is not formulated in such strong terms:

"We believe that... pitch movements that are interpreted as relevant by the listener are related to corresponding activities on the part of the speaker. These are assumed to be characterized by discrete commands to the

---

[25] this has not prevented some analysts from using the topline as a reference line, for instance, Cooper and Sorensen, 1981.

ach eachvocal cords and should be recoverable as so many discrete events in the resulting pitch contour, which may present themselves at first sight as continuous variations in time."

(Cohen and 't Hart 1967, 177-8, quoted in 't Hart et al. 1990, p.39). The research group explicitly denies such a link: "...we do not claim that the listener must have direct knowledge about the laryngeal activity on the part of the speaker"('t Hart et al. 1990, p.40). However, there is evidence to suggest that such a link is unavoidable when it comes to the question of declination. This point will be addressed shortly.

There are two main stages in this 'distillation' process. Firstly, 'close copy stylisations' of F0 contours are made, and constructed in such a way that they are considered perceptually identical by subjects in a perceptual test. The close-copy stylisations are made by fitting straight line sections (in the log frequency domain) to the natural F0 contour, in the case of each such contour, as few as are necessary for subjects to make the judgment of perceptual identity[26]. This process removes a lot of the variation due to segmental coarticulation effects from the F0 contour (though not all; Silverman 1987 suggests that many such effects, notably that of intrinsic pitch, are perfectly detectable by the average listener).

Secondly, standardised pitch movements are chosen to model the intonation contours in a corpus of the language. This is done firstly by determining which contours are 'perceptually equivalent'[27] when constructed from the kinds of straight line segments used in the first stage of modelling. Then, amongst each group of such contours, averages of sizes, durations and positions of the pitch movements are taken. An iterative process of adjustment and analysis-by-synthesis (concluded by a formal acceptability test) is then undertaken to adjust these values to a standardised set of pitch

---

[26] 'Perceptual equality' is the term used in the Eindhoven School's publications. It refers to identity of perception elicited from objectively different intonation contours, in which subjects in a psychophonetic task (who would not thus have to be native speakers of the language being studied to perform the task) could not distinguish between them.

[27] 'perceptual equivalence' refers to the equivalence of contours which may not be 'perceptually equal' - that is, could be detected by subjects in psychophonetic experiments as distinct - but would be accepted as instances of the same contour, imparting the same message, by speakers of the language.

movements which conform to a single system of parameters. In the case of Dutch, the following pitch movements and other parameters were eventually selected to represent the building blocks of intonation contours:

/ Full Rise (of 6 semitones @ 50 semitones/sec., thus over 120 ms)

/ Half Rise (of 3 semitones @ 50 semitones/sec., thus over 60 ms)

\ Full Fall (of 6 semitones @ 50 semitones/sec., thus over 120 ms)

\ Half Fall (of 3 semitones @ 50 semitones/sec., thus over 60 ms)

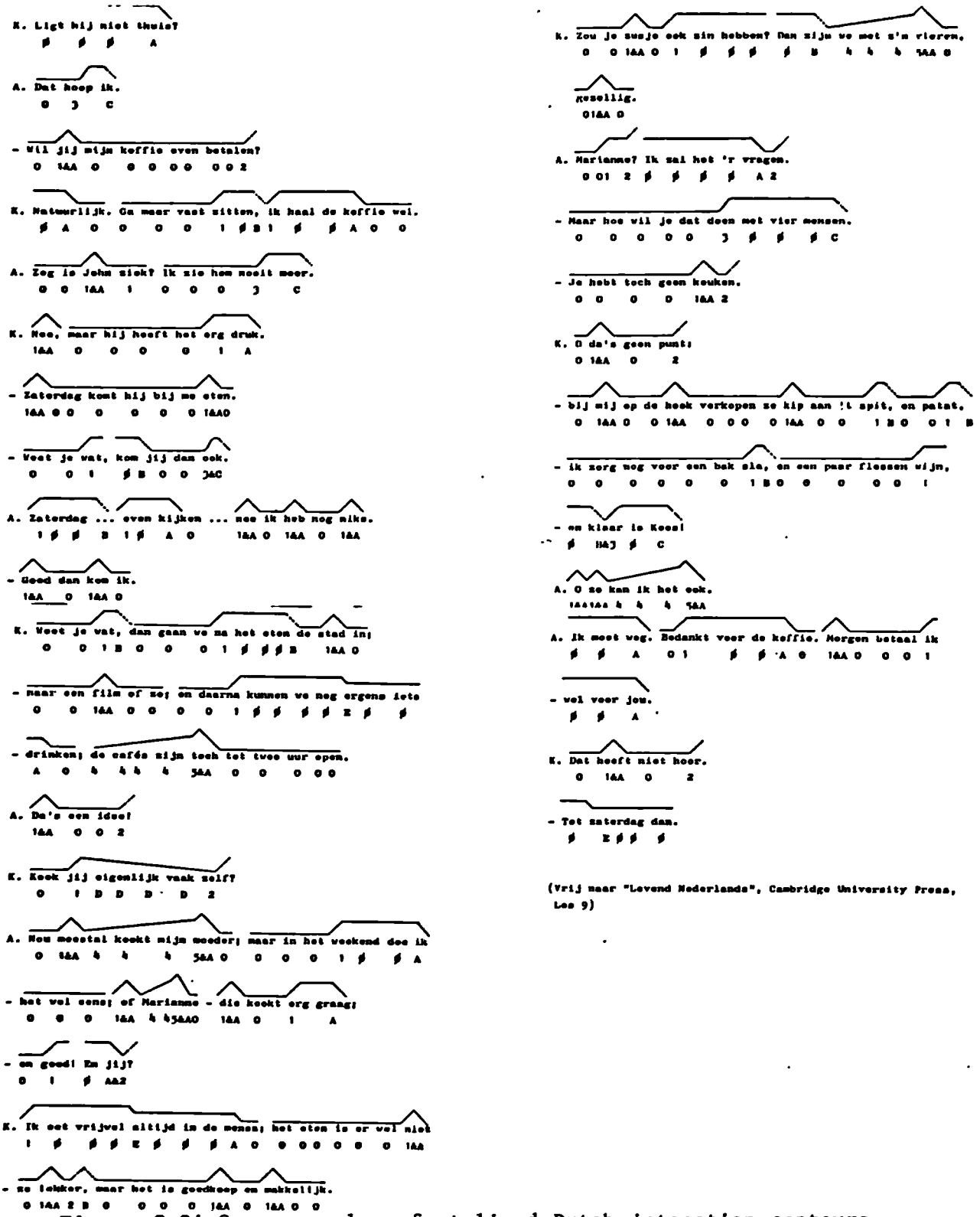
Table 2.2 The pitch movements of Dutch
(taken from 't Hart et al's Table 4.1, 1990 p.73)

	Transcription symbol:									
	1	2	3	4	5	A	B	C	D	E
Direction										
rise	x	x	x	x	x					
fall						x	x	x	x	x
Timing										
early	x				x		x			x
late			x			x				
very late		x						x		
Rate of change										
fast	x	x	x		x	x	x	x		x
slow				x					x	
Size										
full	x	x	x	x		x	x	x	x	
half					x					x

The course of these falls and rises is between declining reference lines. Thus, although for the most part, their treatment considers only a declining topline and baseline, it implies also a declining midline, to and from which the half-falls fall, and to which the half-rises rise.

The pitch movements are further categorised into forms which depend on whether they occur early, late or very late relative to the onset of the vocalic nucleus of the syllable with which they are associated. This categorisation yields five falls (lettered A-E) and five rises (numbered 1-5), as in Table 2.2[28]

[28]There are further groupings of these pitch movements into configurations, which are 'molecules' of co-occurring pitch movements; these are then built into intonation contours, according to the grammar mentioned above. These intonation contours can also be grouped according to the specification of a core set of 'root' configurations (there are also 'prefix' and 'suffix' configurations) into intonation patterns. These higher level



(Vrij naar "Levend Nederlands", Cambridge University Press, Les 9)

Figure 2.24 Some examples of stylised Dutch intonation contours, taken from Cursus Nederlandse Intonatie, Collier and 't Hart 1981.

---

groupings are not of concern here.

Some examples of contours generated using these pitch movements appear in Fig. 2.24, taken from 'Cursus Nederlandse Intonatie', Collier and t'Hart, 1981. These contours have been shown without any applied declination. When they are fully generated, the pitch contour is computed with a downward tilt on the frequency axis, which is scaled logarithmically, in semitones. The slope that is used (in semitones per second) is

$$-11/(t+1.5)$$

(where t is the duration of the utterance in seconds) for utterances that are less than or equal to 5 seconds in duration, and

$$-8.5/t$$

for longer utterances. The baseline is computed with a fixed end point, and the topline simply runs parallel to it (at a distance of 6 semitones), so that the longer the utterance the shallower the slope; but the effect of the adjustment of 1.5 in the divisor has the additional effect of increasing the starting pitch of the declination line with increasing duration (up to 5 seconds).

2.2.4.1.2 The Eindhoven school analysis of English

The approach to modelling the intonation of English, as detailed in De Pijper (1983), is similar to that used in modelling Dutch. Again, close-copy stylisation was used to derive minimal straight-line segment versions of natural F0 contours, and an iterative analysis-by-synthesis method used to determine standardised pitch movements from which pitch movements could be constructed. However, De Pijper expedited that second task by availing himself of the inventory of contours that Halliday specified for English, in Halliday (1970), as the set of groups each one of which was used as a target in testing for perceptual equivalence. The resulting set of pitch movements was as follows:

Table 2.3 - The pitch movements of English
(taken from De Pijper, 1983, Table 3.2, p.49)

		Steep		Gradual	
		Half	Full	Half	Full
Rise	Early	1121	1141		
	Middle	1122	1142	1220	1240
	Late	1123	1143		
Fall	Early	2121	2141		
	Middle	2122	2142	2220	2240
	Late	2123	2143		

(Note, the numbers follow the following code :

" A denotes direction: 1 rise,
2 fall;
B denotes steepness: 1 steep,
2 gradual;
C denotes range: 2 two quarters, or half the range,
4 four quarters, or the full range;
D denotes position: 1 early,
2 middle,
3 late.
", De Pijper 1983, p.48).

In this analysis of English intonation, a more explicit use is made of the middle reference line, as many more pitch movements make use of it than the two in Dutch. Some contours generated by his algorithm appear in Fig. 2.25.

Again, no declination has been applied in the diagrams. The declination slope used in this case is

for t <= 4.82 s. D = -1/(0.13 + 0.09(t))

for t > 4.82 s. D = -1/0.117 (t)

It's application in the shorter utterance of Fig. 2.25 can be seen in Fig. 2.26

2.2.4.1.3 Discussion

Having looked briefly at the Eindhoven school's analysis of Dutch and English intonation, it is now possible to see how the contributions to the downward trend of intonation contours analysed in section 2.2.3 as local phenomena are

incorporated within a global declination function. Firstly, the pitch movements and topline, baseline and midline are all computed on the semitone scale. This naturally accounts for some of the local variability due to Initial Raising, Downstep and Final Lowering on a linear frequency scale[29].

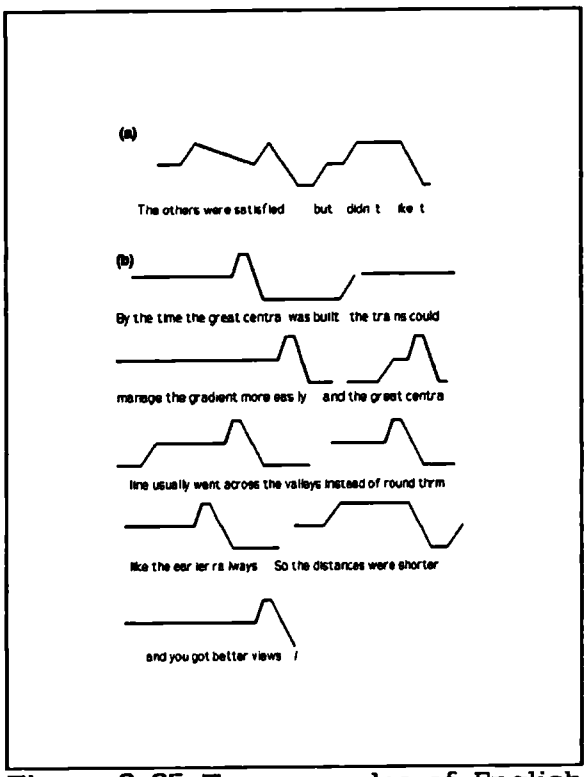


Figure 2.25 Two examples of English contours modelled according to De Pijper's (1983) specifications. (a) Two tone-unit utterance. (b) 'spontaneous dialogue' from Halliday's (1970) intonation course.

Secondly, the process of fitting contours to original ones using standardised pitch movements naturally has the consequence of ignoring some of the local variation caused by Initial Raising and Final Lowering. Thirdly, if Dutch alone is considered (and this is naturally often the case in discussions by and of the Eindhoven School) there is less of a case for arguing the existence of Initial Raising and Final Lowering, because the pitch range used in Dutch is quite small, so the relevant local excursions occur over proportionally less of the pitch range than in languages such as German and English. Indeed, quite a resourceful case had been made by one member of the school against a suggestion that Dutch listeners take Final Lowering into account in assigning prominence to accented syllables (Terken 1989a, commenting on Gussenhoven and Rietveld 1988; also see Gussenhoven and Rietveld 1989).

[29] For the ratio of Hertz to semitones is greater than unity even at the lowest possible frequency of vocal-fold vibration, and so any downward trend attributable to final lowering on a Hz. scale would already be attenuated on a semitone scale. As that ratio increases upon increase in frequency, the attenuation of downward trends due to downstep and initial raising is even greater.

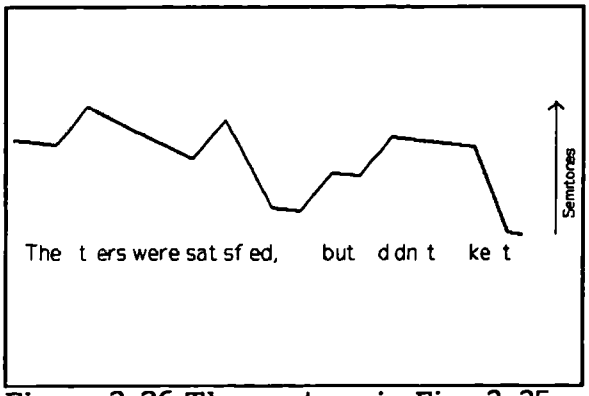


Figure 2.26 The contour in Fig. 2.25a, with declination applied.

Thus far, it has only been seen that the variation attributable to Initial Raising and Final Lowering could be included within a global declination function. What of downstep? Well, in the treatments of both Dutch and English, a step down from a high pitch to a mid pitch can be accounted for within the treatments by using a half-fall. In many cases of downstep, then, it would remain a local phenomenon, and not be incorporated within the global declination function. The situation is less clear with downstep sequences. 't Hart et al. suggest (1990, p.159, para. 5) that downstep iterations are possible in Dutch, but it is difficult to see how they could be computed using their model without requiring associated increases in the pitch range. For each downstep requires a downward excursion of 3 semitones in their Dutch intonation model, so two downsteps covers the whole of the pitch range. If the pitch range were expanded, it would not be an isolated requirement; the half rise '5' appears to be able to start <u>from</u> the topline after a gradual full rise '4' or steep full rise '1'. But even then, there would be a possible problem: downstep sequences are generally taken to follow an exponential line of descent (indeed, Silverman 1987 accounts wholly for downstep in English by scaling his pitch targets on an exponential pitch scale[30]). Yet the semitone scale does not reflect such exponential scaling. In fact, consecutive equal steps down of significant magnitude on the semitone scale represent steps

[30] This approach to scaling targets in intonation contours is to take Pierrehumbert's requirement that downstep is pivotal in the generation of intonation contours (see chapter 3) to an extreme. It is adopted for reasons of computational elegance, and rather passes the buck on the question of accounting for the effect of downstep on the scaling of prominence, for which task it requires much experimental corroboration to sustain the suggestion that an exponential scale is <u>in general</u> appropriate.

down of an increasing interval on most of an exponential pitch scale[31] with a reasonably low asymptote. So the sequences generated in Dutch (within a possibly greatly expanded pitch scale) would be unlikely to represent precisely the downstep sequences found in natural language. As for English, the same possibilities and problems arise.

Another possibility for the Endhoven School would be to extend their standardisation strategy; and to let the declining midline model any subsequent downsteps, which would have relatively small intervals in any case. It is not possible to presume that this would be the best thing to do in the absence of hard data on downstep sequences in Dutch, but it could perhaps be made to adequately account for downstep sequences in English. If that were so, then only the first step-down of a downstep sequence (or an isolated such step-down) would remain a local phenomenon contributing to the downward trend in intonation contours, and this would have the same

---

[31] For example, Silverman's scale is defined by the equation

$$F0 = FLOOR + UNIT * (LOGBASE^{T})$$

where F0 are units in Hz, FLOOR is the bottom of the speakers range, UNIT is the difference between a reference level REF and FLOOR in Hz. and LOGBASE is the ratio of the whole range to UNIT (i.e. the reciprocal of the proportion of the range at which REF is fixed), and T is the prominence (positive above REF, negative below it). Given the following values:

FLOOR = 75Hz
UNIT = 15Hz
LOGBASE=4

then for a value of T=1, $F0 = 75+15*4^{1} = 135Hz$,
and for a value of T=0.5, $F0 = 75+15*4^{0.5} = 105Hz$
and for a value of T=0, $F0 = 75+15*4^{0} = 90Hz$.
If the distance between values of 135Hz and 105Hz is computed in semitones, the value (according to 't Hart et al's formula - 1990, p.24 - in which

$$D = 12/\log(2).\log(f1/f2),$$

where D = distance in semitones, f1 and f2 are the respective linear frequency values) comes out at 4.35 semitones. Using the inverse function of that expressed in 't Hart et al's formula, a further step down of the same value in semitones yields a linear frequency value of 82Hz. For a similar second step down on the linear frequency scale, then, a second step which is larger than the first is required on Silverman's exponential scale (i.e. $0.5-\log((F0-FLOOR)/UNIT)/\log(LOGBASE) = 0.5-\log((82-75)/15)/\log(4) = 1.05$; since $x^{y} = e^{y \cdot \log(x)}$, and so $LOGBASE^{T} = \exp(T.\log(LOGBASE))$. Note that this relationship between the semitone and Silverman's exponential scale only holds for step values which are sufficiently large not to be approaching the limit of linearity on a quantized exponential scale of reasonable precision, and a fortiori, which are as large as are involved in downstep sequences; it also only holds generally for the exponential scale above the value REF.

status as all non-gradual falls, its contribution to the downward trend being considered naturally excluded from a declination function[32].

This approach to modelling the downward trend in intonation contours treats declination as a 'frame of reference', to use Ladd's (1984) parlance, within which local variation is scaled. As it has been presented here, the Eindhoven approach seems to require that parts of the intonation contour actually touch or be comprised by sections of the declining topline and baseline (and midline), because of the process of standardisation of pitch movements, which assumes the physical existence of the reference lines. However, that process is just part of a heuristic method, and they would have to accept that when presented with an utterance in which there were no necessary physical contact between its intonation contour and the putative declining reference lines, a listener would use some internal computation of the latter in order to scale appropriately his/her perception of the intonation. This, at least, is what is implied by the suggestion by the Eindhoven school that declination is an ever-present concomitant of intonation contours, regardless of the physical excursions of F0, underpinned primarily by declining subglottal pressure (P_s)(which accounts for a particular decline in F0 resulting from the involuntary F0/P_s ratio)('t Hart et al. 1990 Ch.5)[33]. To the extent that this is so, this method of determining a declination line effectively makes declination equivalent to the transform of a coordinate system in which pitch points are plotted.

A little must be said here about the extent to which that is so. The approach of the Eindhoven school is notably consistent with many American approaches to modelling intonation, which, as we have noted earlier, are oriented towards the productive rather than the perceptual aspects of speech communication. This is explicitly acknowledged by 't Hart et al :
"..only those F0 changes are relevant for perception that have been intentionally produced by the speaker as physical properties that are cues

---

[32] The case of the gradual fall has not been touched on thus far. Further consideration of the phenomenon is given in Chapter 5.

[33] It should be pointed out, however, that they do observe that Leroy (1984) has discovered that physical changes in the F0 contour can disrupt the listener's mental projection of the declining baseline. More of that in Chapter 4.

to the intonation pattern that he wants to produce. Then, for an F0 change to be detected as such a cue, it should not only be above some psychophysical threshold, but at the same time be recognised as the result of some purposeful action on the speaker's side. This distinct, 'phonetic', mode of pitch perception in speech echoes a similar claim for segmental perception, embodied in the revised 'motor theory', as formulated by A. Liberman and Mattingly (1985:2): 'The first claim of the motor theory, as revised, is that the objects of speech perception are intended phonetic gestures of the speaker'. " (1990, p.70).

The course of declining subglottal pressure, which might be considered to be an automatic consequence of the physiological mechanisms involved in producing speech, is equally as valid, as an F0 change 'relevant for perception' 'intentionally produced by the speaker', as local accentual excursions, in the current stated position of the Eindhoven school. This is because a speaker is understood by them to choose between at least two control strategies in producing speech: one for speech, which involves declining subglottal pressure of a particular slope and drop, depending on the length of utterance, and one for song, chant or any situation in which sustained phonation at a particular level is required, which involves the maintenance of a particular level of subglottal pressure (see Johnson and Grice, 1990, for a discussion of chant and stylisation). Each of these control strategies involves the choice by the speaker of an appropriate physiological setting at the start of the utterance, whereupon no further monitoring is required (the control mechanism is open-loop). This initial choice means that the F0 change of declination could be considered 'intentionally produced'. Thus, if the declination line retrieved by perceptual investigation of the 'declination effect' were the same as that determined as attributable to subglottal pressure decline, then there would be additional conformity with the 'motor theory' of speech perception.

The way standardised pitch lines are fitted to F0 contours reflects this two-level conformity with the motor theory. Lines are fit to local accentuation configurations on the basis of acceptable 'phonetic' fit to the original F0 contour; this criterion uses the 'close copy' stylisation as a starting point, which requires a fit to the gross shape of the F0 contour. Thus the actual F0 values are used as a partial reference for determining the standardised

contours' fit. In the case of the declining reference lines, there is a preconception about the kind of lines which ought to be made to fit. There is considerable scope for fitting a variety of reference lines, as the amount of physical contact between the reference lines and the F0 contour is not stipulated[34]. What is missing is a criterion for determining what of the F0 contour needs to be actually made to physically fit along the declining reference lines. In short, a set of criteria for admissible abstraction is required.

A case in which declination is more clearly treated as an abstract coordinate system is found in Pierrehumbert's (1980) approach. There will be an extensive dicussion in the next chapter of the contribution of downstep to downward trends in her thesis. It is pointed out there that the phonetic units in which her contours are computed are baseline units which mediate Pierrehumbert's global declination component. The following gives some detail of this transform of fundamental frequency.

2.2.4.2 Pierrehumbert's 1980 account of declination

For Pierrehumbert, declination "is a gradual downdrift and narrowing of the pitch range, which occurs within the body of the intonation phrase and frequently over the course of several intonation phrases" (1980, p.116). In linear frequency terms, the same is true of the pitch range described by the Eindhoven school's declining topline and baseline : parallel in the semitone scale, they converge from left to right (i.e. from t=m to t = m+n, n>0) when the left hand frequency axis has a scale in Hz. However, Pierrehumbert doesn't require an explicit topline to define the upper limit of the pitch range (we have already noted that the Eindhoven's analysis of Dutch intonation includes tones which can in any case transgress the upper limit of the pitch range); the range can be shown to narrow as well as drift down by the scaling of H* pitch accents within the transformed coordinate space defined by the equation

$$P = (p-b)/b$$

where P is the pitch or phonetic value of such a H* pitch accent, p is its value in Hz, and b is the value in Hz of a hypothetical baseline which declines

[34]They claim that, in principle, there need be no such contact. However, in practice, there is always such contact.

linearly from the start of an utterance to its end. To show this downdrifting and narrowing effect, it is enough to consider the following case: if the baseline is taken as starting at 100Hz and ending at 75Hz, then a H* pitch accent scaled at its start would have a value of 150Hz for a phonetic value of 0.5 baseline units (0.5 = (150-100)/100). A H* pitch accent scaled at its end would have a value of 112.5Hz for the same phonetic value of 0.5 baseline units (0.5 = (112.5-75)/75). This demonstrates the downdrifting effect. The narrowing effect can be seen from the fact that in the case of the first accent, the distance between the value for the H* pitch accent and the baseline is 50Hz, whereas in the second, it is 37.5 Hz.

It is important to note that the baseline is a hypothetical one. That is, there is in theory no part of any pitch contour which needs to descend to it or follow it (and this is more clearly the case than in the Eindhoven School's model) - it acts rather as a reference line for scaling pitch accents (and in fact, by extrapolation, all pitch values in a contour). Pierrehumbert determines the start and end values for the variable b from a production experiment involving the scaling of the pitch accents involved respectively in background and answer information in different contexts[35]. She then uses these as the values appropriate for all utterances by a particular speaker. Futhermore, she postulates that there is a constant amount of declination in each utterance; that the baseline has constant drop. These extrapolations from a single experimental situation rest on the assumption that declination is an invariant backdrop to all utterances, with presumably physiological causes.

The use of such an abstract declination line suggests that a listener, in scaling the prominence of accents in a person's voice, can't rely on F0 values in the intonation contour as cues for determining the declining baseline of

---

[35] The values for b are chosen as best fits for parameters (b1 and b2, the value of the baseline at the position of the pitch accents) in an analysis involving linear regression of the value in Hz of the second accent on the height of the first over a number of utterances; the other parameter is c, the constant expressing the relationship between the phonetic value of the answering accent and the background accent in the formula

$$A = cB$$

where A and B are the phonetic values of the answering accent and background accent respectively. The values for b are thus not taken from physical baseline values in the F0 contour.

that speaker. Instead, they would have to rely on such cues as voice quality, or filter quality of the oral and nasal tracts as imparted by formant values, allied with knowledge about the correlation between vowel space and larynx size, whence likely pitch values (Pierrehumbert herself points this out - 1980 p. 136). In a model employing similar abstraction of the declining baseline, but without the claim that each speaker has their own particular baseline drop, a listener might have a mental representation of a normalised declining baseline which they use to decode from the pitch values on accented syllables (themselves normalised by some process) the prominence the speaker intended to impart (always assuming that downstep or some other local adjustment function is not a conflating factor). These possibilities complicate the experimental procedure for perceptual testing of the 'declination effect' and thence the perceived declination line, because they require controllable quantitative manipulation of at least (a) voice quality, (b) formant values according to some hypothesized relationship with fundamental frequency, and (c) listeners' expectations of the normal behaviour of intonation contours, and the extent to which controlled variation of F0 in an experimental presentation is likely to conflict with such expectations. Of these, although the first two have their own difficulties, the last seems impossible to overcome, and this is one of the reasons that the experimental procedures adopted in Chapter 6 augment perceptual tests with tests of perception in production.

2.2.4.3 Liberman and Pierrehumbert - Doing away with declination altogether

In Pierrehumbert's 1980 thesis, the declination function accounts only for a part of the downtrend in an intonation contour. In work she did after that thesis, in collaboration with M. Liberman, reported most fully in Liberman and Pierrehumbert (1984), the downtrend is not considered to be a phenomenon that is necessarily always present in an intonation contour (although they do not rule out that possibility). Instead, particular aspects of downtrends that do exist in specific intonation contours are attributed to the following: (i) downstep, (ii) final lowering, (iii) initial raising, (for all three see section 2.2.3) and (iv) changes in pitch range, reflected in initial peak height. There is no separate global declination function, although, again, the possibility that one exists is not ruled out. Moreover, the declination function of Pierrehumbert,1980, is explicitly discarded as not

being capable of accounting for downstep sequences in different pitch ranges with the required accuracy (Liberman and Pierrehumbert 1984, p.210).

Liberman and Pierrehumbert's model is as follows (ibid., p. 192) :

(a) General F0 Transform

$$T(P) = P - r$$

where P = the peak target value on an accented syllable in Hz., and
r = a reference value in Hz.

(b) Downstep

$$T(P_a) = s \cdot T(P_{a-1})$$

where P_a is the peak target value on the ath step-accented syllable in Hz.

(c) Answer-Background relation

$$T(P_A) = k \cdot T(P_B)$$

where P_A is the peak target value in Hz. of the 'answering' accent, and P_B is the peak target value in Hz. of the 'background' accent.

(d) Relation of r to initial accent target

$$r = f \cdot (P_0 - b)^e + d + b$$

where P_0 = the peak target value of the first accented syllable in Hz, b = the absolute baseline in Hz. for a speaker, d = the minimum value above b that r may be at, f is an arbitrary constant used to scale e, an empirical exponent.

(e) Final Lowering

$$P \rightarrow r + l \cdot (P-r) \;/\_\; \$$$

where $l<1$, P is the peak target value of an accented syllable in Hz., '/' refers to context, and '$' indicates a phrase boundary.

These rules were devised to account for the intonation in a variety of contexts, though they were developed on the basis of two specific sorts of contour: a downstep sequence, and a pattern exhibiting the answer-background relationship, as was discussed in section 2.2.4.2 above in connection with Pierrehumbert's 1980 account of declination. The relationships expressed in rules (b) and (c) are obviously specific to their

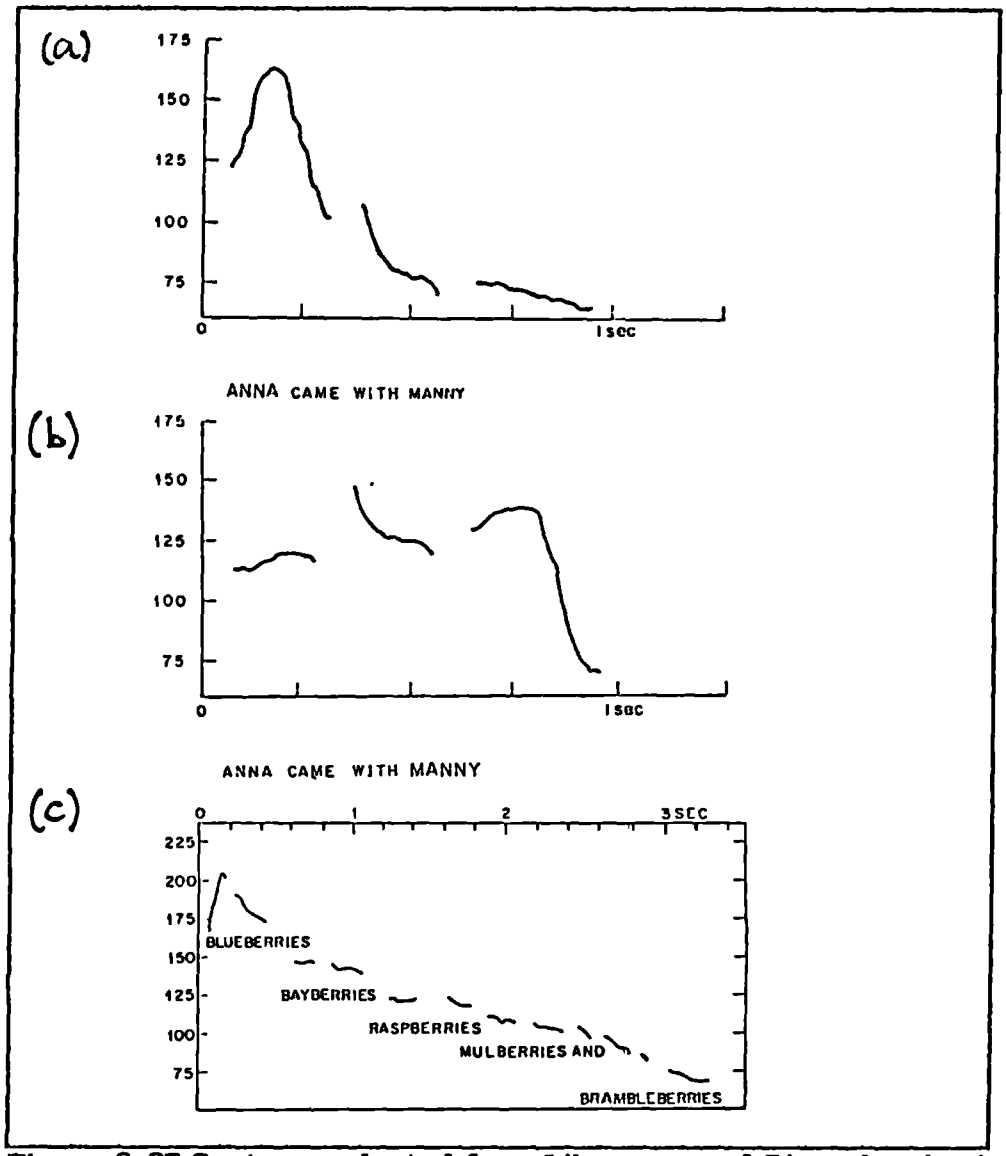


Figure 2.27 Contours adapted from Liberman and Pierrehumbert (1984). (a) Answer-Background contour (A-B order); (b) Answer-Background contour (B-A order); (c) Downstepping sequence

contexts. The scaling expressed in rule (e) is common to both contexts, and naturally extends to all utterances which have a terminal fall. It is less clear that it extends to utterances which have, say, a final accent which is purely rising (such as in a L* H H% sequence - see Ch. 3 for interpretation of these symbols). The relationships expressed in rules (a) and (d) relate to the variation in scaling of accent peaks as the pitch range in which they are uttered varies. It is taken that the F0 value at the peak of the first pitch

accent in the contexts studied reflects the choice of the width of the pitch range over the whole intonation phrase[36].

Although this model 'works' in the types of contour tested - and apparently works best, out of a number of competing models, as Liberman and Pierrehumbert demonstrate - there is reason to suppose that the complete removal of a declination function that it presupposes would lead to difficulties in accounting for contours of many another sort. The directly modelled contours appear in Figs. 2.27a-c.

In none of these contours is there what could count as a fairly long unaccented stretch of speech. In Fig. 2.27a, the words 'came with' are unaccented, but it is difficult to disentangle any trend in the F0 on those words from segmental coarticulation effects; the local interaccentual intonation is similar in Fig.2.27b. In Fig. 2.27c, each of the 'berry' words has a step accent on its first syllable, and the stretch of unaccented material that follows each such syllable is fairly level, as befits the stepping sequence, but again, the effects of segmental coarticulation are not so clear.

In Fig.2.28 is a contour in which there <u>is</u> such an interaccentual sequence. Here, the overall trend in the unaccented sequence is much clearer, and is seen to be gently declining. If the two accents were considered to be H* accents, according to Pierrehumbert's taxonomy, which is what they could be expected to be (within the Liberman-Pierrehumbert paradigm) in the absence of downstep on the second accented syllable, then nothing in Liberman and Pierrehumbert's 1984 rules could account for such an interaccentual decline[37]. Either an alteration to Pierrehumbert's 1980 interpolation rule between H accents would be required to account for it, or a reinstated declination function of some sort[38].

---

[36] As will be seen in Chapter 3, Ladd (1990) has had cause to take issue with that suggestion.

[37] The requirement of separate declination and downstep components on these grounds has been pointed out by Ladd (1983, pp. 48-9, and 1984, pp. 64-5).

[38] It has to be said that that function might also account for some of what decline <u>can</u> be detected on the 'berries' words in Fig. 2.27c and the words 'came with' in Figs. 2.27a and 2.27b.

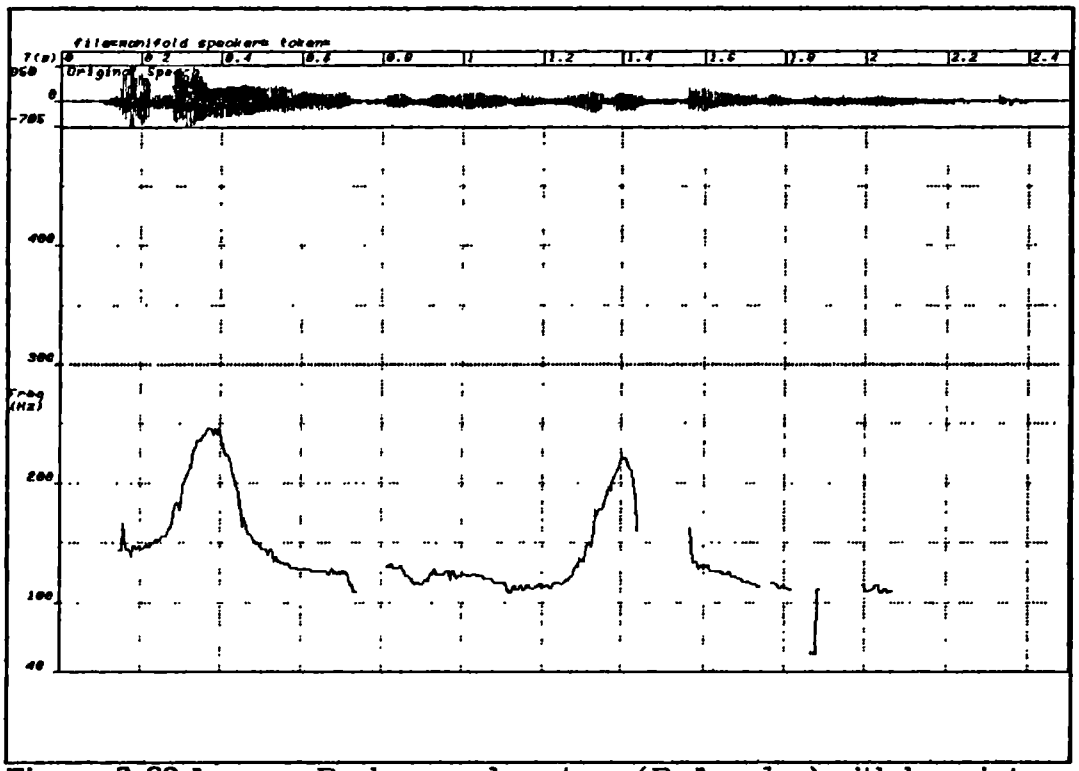


Figure 2.28 Answer-Background contour (B-A order) with long inter-accentual stretch. 'Anna arrived with a yellow manifold in her hand'.

2.2.4.4 A local declination function and a global lowering function

Such a function could be explicitly local, accounting for declination only in such contexts as that just illustrated, where it is clear that there is a declining trend in which accentuation and boundary effects are not directly involved, or it could be a global function, perhaps not having the effect of lowering and narrowing the efective pitch range over the course of a prosodic domain, as Pierrehumbert's 1980 function did, but merely of lowering it.

The former strategy, of implementing local declination functions, is known by the current author to exist in the intonation modules of two text-to-speech synthesis systems - that of Mattingly 1966, and of Johnson and House (1986) (also House and Johnson 1987, House, Johnson 1989, Johnson 1990). In the latter case, all constraints on the relative F0 values on accented syllables were implemented by local first-order probabilistic rules, and declination was implemented as a slow decay of F0 on stretches marked as [level] (either explicitly or as the result of rule operation) in the text. In this case, within a tone-unit (which comprised a sequence of accent-units)

no intonation rules with global operation were present, and the algorithm generated probabilistically varying intonation contours that were for the most part acceptable, although there were a small number of less than well-formed contours that could be synthesised.

The latter strategy, of implementing a lowering (but not by itself narrowing) global declination function of very gradual slope, is the option adopted by Beckman and Pierrehumbert (1986) and Pierrehumbert and Beckman (1988) in their analysis of Japanese intonation (which approach they also recommend extending to English intonation) and has also been suggested as a possibility, as will be seen in Chapter 3, in Ladd's 1990 treatment of English intonation. In the former, an appropriate value for the slope of declination for a particular speaker is derived from analysis of the residuals following linear regression of the F0 of the first accent peak in a two-accent intonation phrase on that of the second. In some speakers there is no identifiable pattern in the mean residuals taken over utterances of increasing length, where the increase in length is due to an expandible interaccentual stretch. In others, however, the residuals show a declining trend, with the values in shorter utterances being positive and those in longer utterances negative. This means that the difference between the peak on the second accented syllable and that on the first is greater in the case of the longer utterances than in the case of the shorter utterances. So, the longer an utterance is, the more there is a tendency for a second accented syllable to be lower in pitch. This argues for a time-dependent declination function, although, as Pierrehumbert and Beckman observe (1988, p.20), there is nothing in the data to choose between a linear or non-linear function. The slope for one particular individual was 10 Hz. per sec. This value could thus be used to construct a shallowly declining baseline (leading back from a fixed final low L% value) relative to which intonation contours are computed. It is important to note that this global time-dependent declination function is different from that in Pierrehumbert 1980, which is time-independent in respect of drop; that is, which has a fixed drop but a slope that varies with time. At the same time, there is no narrowing of the pitch range inherent in the declination function; all the narrowing is performed by a catathesis (downstep) rule.

2.2.4.5 A componential approach - analysis of Danish by Thorsen

In the studies discussed so far, the implication has always been that any global declination function would be fixed for a given duration of prosodic unit, regardless of the utterance, with some variation possible across speakers. This is in keeping with the suggestion that declination is

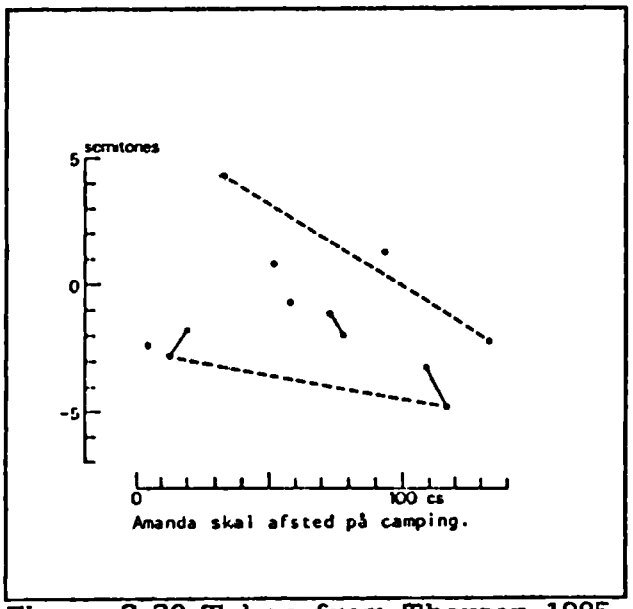


Figure 2.30 Taken from Thorsen 1985, Fig. 2. Salient pitch points averaged over three instances of the utterance "Amanda skal afsted pa camping". Dashed lines are declining topline and baseline.

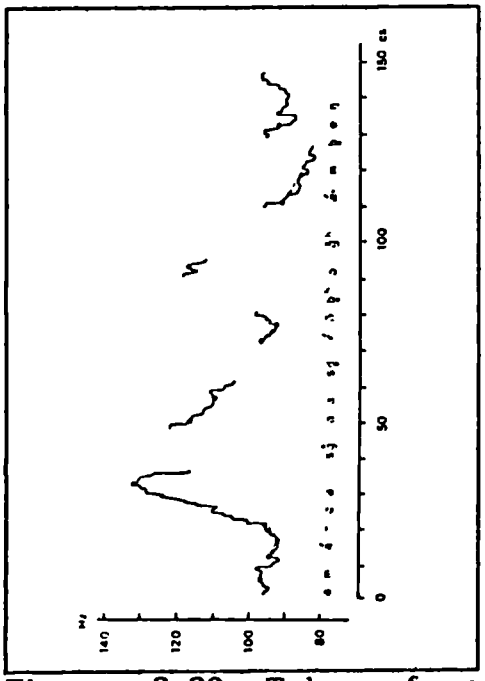


Figure 2.29 Taken from Thorsen 1985, Fig.1

an automatic byproduct of the speech-production mechanism, not varying under any sort of phonological controls; and that there need be no cues to a declining reference line or framework within the F0 signal, though they could be computed in such a way as to be consistent with that signal, from cues in separate channel(s) of the speech signal (such as voice quality) or on the basis of more abstract time-dependent expectations. An alternative view is put forward by Thorsen, in her studies of Danish intonation (Thorsen 1978/9, 1980,1983, 1985,1986). For her, an intonation contour (and, a fortiori, one in Danish) comprises accentual components and global phrase level components, both sorts of which are actually physically manifested in the F0 signal. For example, the contour in Fig. 2.29 shows a Danish F0 contour analysed from a declarative utterance of the sentence "Amanda skal afsted pa camping". There is an evident downward trend in the utterance. Lines which physically identify this downward trend can be fitted through the means of points aligned with vowel onsets identified in a number of like utterances, as

in Fig. 2.30. The line through the stressed syllables (whose start and end points are connected by solid lines in Fig. 2.30) marks the 'intonation contour' of an utterance. The pattern of spots after the onset of each stressed syllable marks the stress contour for that stressed syllable. For the dialect that Thorsen largely treats of (Advanced Standard Copenhagen, or ASC, Danish), this consists of an initial high rise in F0 (except when the stress group contains only two stressed syllables, in which case the second syllable is just at the same pitch as the stressed syllable), followed by a sequence of syllables having a falling F0 to a point in the F0 range which may be at, below or above the intonation contour. (For the present discussion, the topline in Fig. 2.30 is of use only in marking the height of the stress accents). There are two other components in the intonation contour, viz. a microprosodic component (which accounts for segmental coarticulation effects), and (in words containing stod) a stod component. No explicit formulas for the generation of intonation contours are provided by Thorsen, but she does say that "the sentence and stress group components may be regarded as simply additive from a productional point of view", (p.1980, p.100) on the assumption that there is a physiological basis for this.

Thorsen's treatment differs from the superficially similar treatments involving scaling of accent configurations within declining tramlines of the Eindhoven School in two ways. Firstly, although the Eindhoven School accepts the distinction between accentual and intonation elements of an F0 contour, it rejects the idea that the two are separate components. That is, it rejects the following hypothesis:

"Hypothesis 1 : Those pitch movements that occur on prominent syllables are entirely and exclusively caused by the accentuation demands of the utterance; the remaining pitch movements (and declination) result from the requirements of intonation proper" which 't Hart et al. gloss as follows: "This hypothesis states that a pitch contour is a sequence of pitch movements that are caused either by the accentual or by the intonational demands of the utterance, but never by both requirements simultaneously". ('t Hart et al. 1990, p.96[39]].

[39] This gloss could be seen as rather misleading, given Thorsen's claim about additivity, but since she provides no quantitative model, her analysis is bound to be susceptible to a variety of interpretations.

It does accept another hypothesis, (see 't Hart et al. 1990, p.97) which states that the intonation of an utterance determines what kinds of pitch movements are used to mark the accented syllables.

Secondly, for Thorsen, different slopes of the intonation component mark different linguistic functions. Fig. 2.31 (taken from Thorsen 1980, fig. 1) summarises her model for short utterances in ASC Danish, and illustrates that a flat intonation component marks a syntactically unmarked question, a steeply falling intonation component marks a declarative statement, and intonation components of intermediate slope mark 'interrogative sentences with word order inversion and/or interrogative particle, and nonfinal periods'[40]. If her approach were applied to English, it would therefore be no problem for her that the overall trend in an F0 contour such as in Fig. 2.2 were rising, because the intonation line on which the 'stress groups' rode would be expected to be rising in a question with word-order inversion, just as it would be expected to be falling in a statement such as in Fig. 2.1.

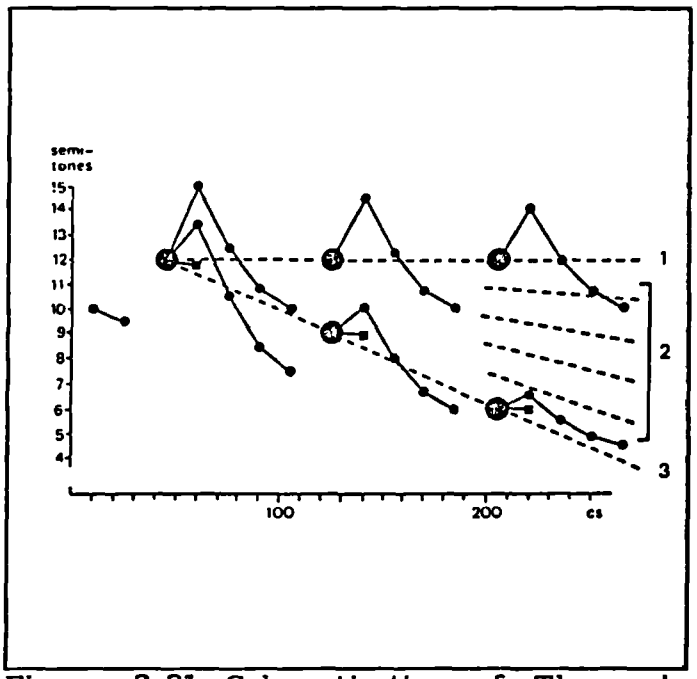


Figure 2.31 Schematisation of Thorsen's componential model of ASC Danish intonation. Dotted lines mark 'intonation' contours for differing sentence types. Full lines mark stress contours.

The approach that Thorsen's analysis exemplifies has become known in recent prosodic research as the 'Contour Interaction' model of intonation, and is one in which the F0 signal is seen to be constructed from a number of different

[40] For longer utterances, there is an additional (again presumably additive) downward moving intonation component, without differences in slope corresponding to linguistic function, which could be taken to have a grouping function (Thorsen 1985,1986).

(usually additive) components, all of which have some physical manifestation in that signal. It contrasts with the 'Tonal Sequence' model of intonation (Nolan 1984, Ladd 1983) in which the F0 signal is seen to comprise a linear sequence of the phonetic exponents of abstract tonal categories, scaled within a framework or coordinate system which is usually tilted to account for declination. Pierrehumbert's (earlier) analyses fall into the latter category, as do Ladd's. The 'contour interaction' approach is reminiscent of the method of Fourier analysis of a signal into component sine functions[41].

2.2.5 Modelling declination as a declining DC component or non-stationary trend; a digression

The analogy drawn at the end of the last section should be justified. When measurements are being performed under Thorsen's method, the intonation component that comprises part of the F0 contour under her analysis is determined by manual means, and could not be determined by an automatic method as simple and consistent as Fourier analysis - at the very least, some rule-based strategy would be needed to identify the stressed syllables through which the intonation line passes. However, there is a sense in which the intonation line so retrieved is acting as some sort of (declining or inclining) DC component in an F0 signal which comprises also an AC component in the form of accent (or stress) contours. It is interesting to speculate on a method whereby a putative such DC component could be retrieved automatically from the F0 signal.

Any signal is considered to consist of at least a DC component, with a possible AC component. A very simple form of signal comprises just a DC level; but any AC signal also has a DC component. This component may be at a level which remains constant, but in many systems, for instance, electrical systems, there can be short-term DC loss. For instance, a battery cell may dissipate energy at a particular rate, such that the DC voltage across the cell terminals is gradually declining. The concept of DC energy level decline is thus one used by electrical engineers, and is readily assimilable into a model

---

[41] Though the word 'interaction' implies non-orthogonality, where orthogonal decomposition is the basis of Fourier analysis, the discussion in Section 2.2.5 postulates that the concept of a separate global intonation component is a cognate of a separable DC component in an arbitrary signal. This is the link between the two different sorts of analysis.

of intonation which treats the F0 contour as comprising a combination of a very slowly varying and more rapidly varying components.

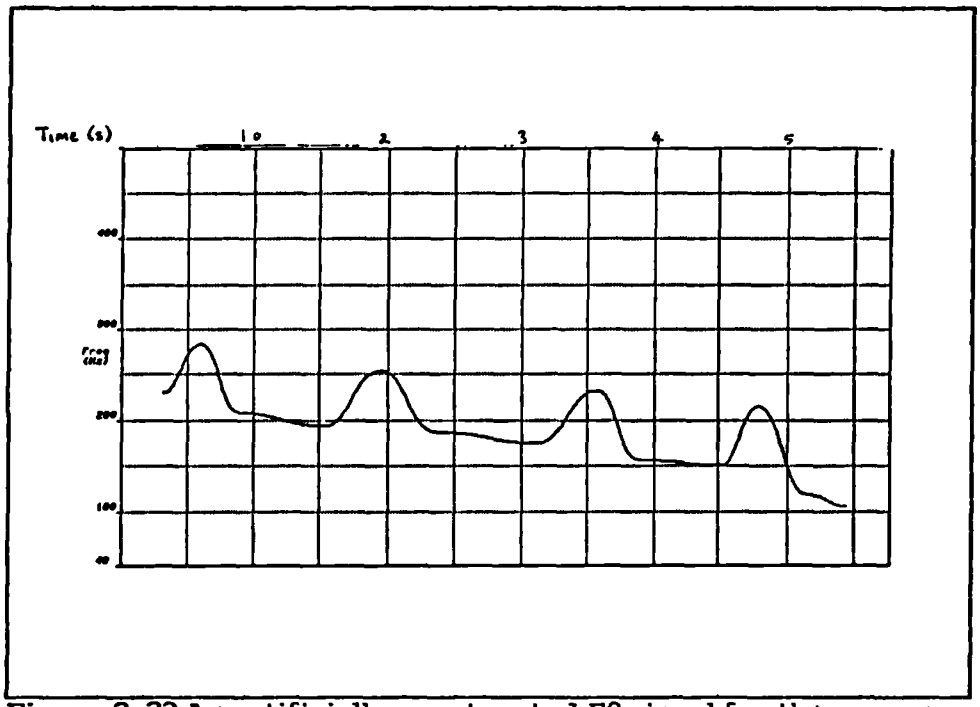


Figure 2.32 An artificially constructed F0 signal for the purposes of illustration. Timescale = 0.5secs at each vertical grid line.

The interpretation that is to be made of a declining DC component in a signal is thus that the energy at zero frequency is constantly declining in that signal. There is, however, a methodological problem associated with that interpretation, which is that it is not possible to retrieve a shifting DC component from an AC signal by standard analytic methods. The DC component has to be considered constant. For example, the DC component of the signal in Fig. 2.32 can be retrieved by the application of a moving-average filter with a very large time window. The large size of the time window is enough to smooth out the high amplitude low-frequency energy which is in the large, relatively slowly moving excursions in such a signal. The output of the filter is a constant value, which to all intents and purposes can be considered the DC level (Fig. 2.33). However, it is clear that there is a declining trend in the signal, and that this ought to be retrievable in some way.

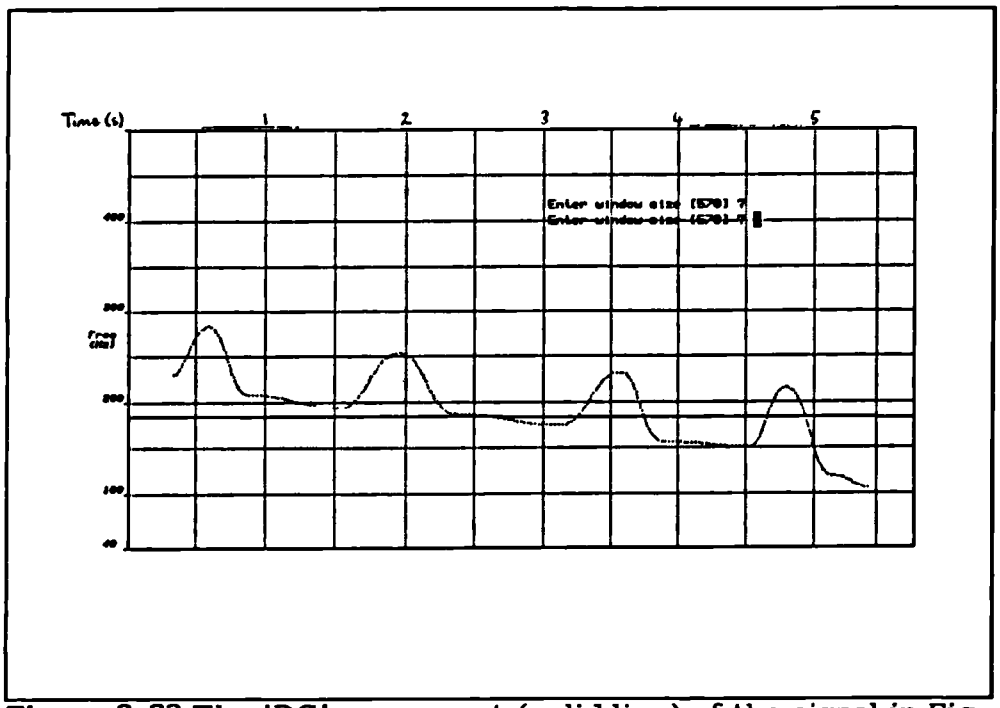


Figure 2.33 The 'DC' component (solid line) of the signal in Fig. 2.32, derived by repeated convolution of a 5700ms wide Hamming window with the input signal.

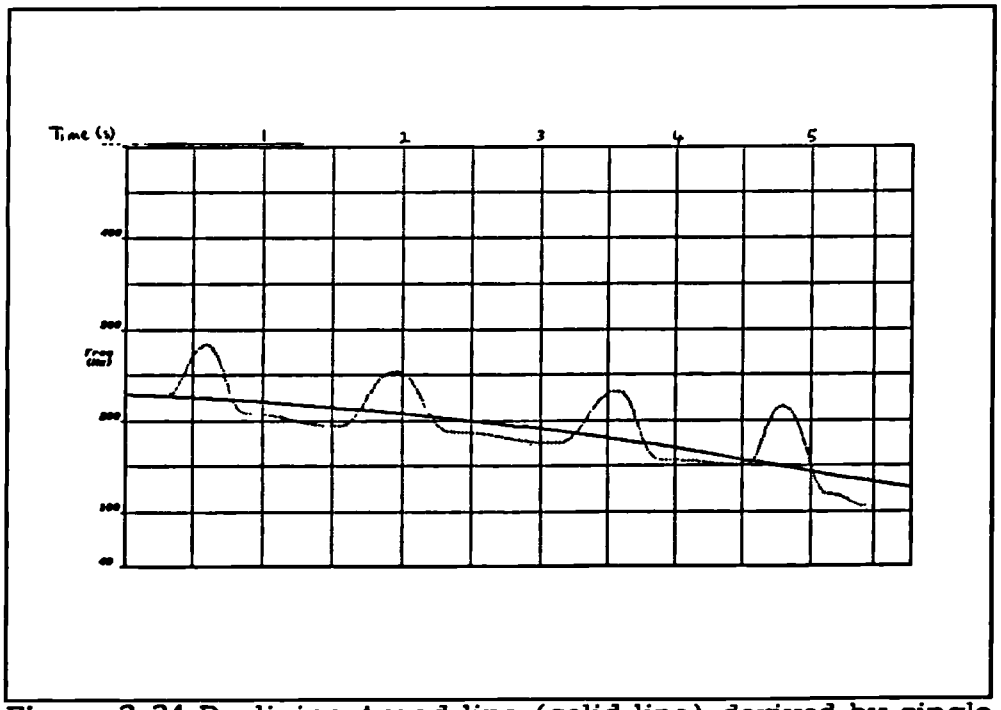


Figure 2.34 Declining trend line (solid line) derived by single instance of convolution of 5700ms. Hamming window with signal in Fig. 2.32.

An approximation to the declining DC level could be made by low-pass filtering using not so great a windowsize as the one which returns the single DC level. The results of such an operation can be seen in the smoothly declining line of Fig. 2.34 (windowsize=570ms, compared to the effective windowsize of over 50 secs. which yielded the DC level in Fig. 2.33). However, it might not be clear what the precise size of the window for use in such an operation should be. A similar signal (Fig. 2.35) has larger local

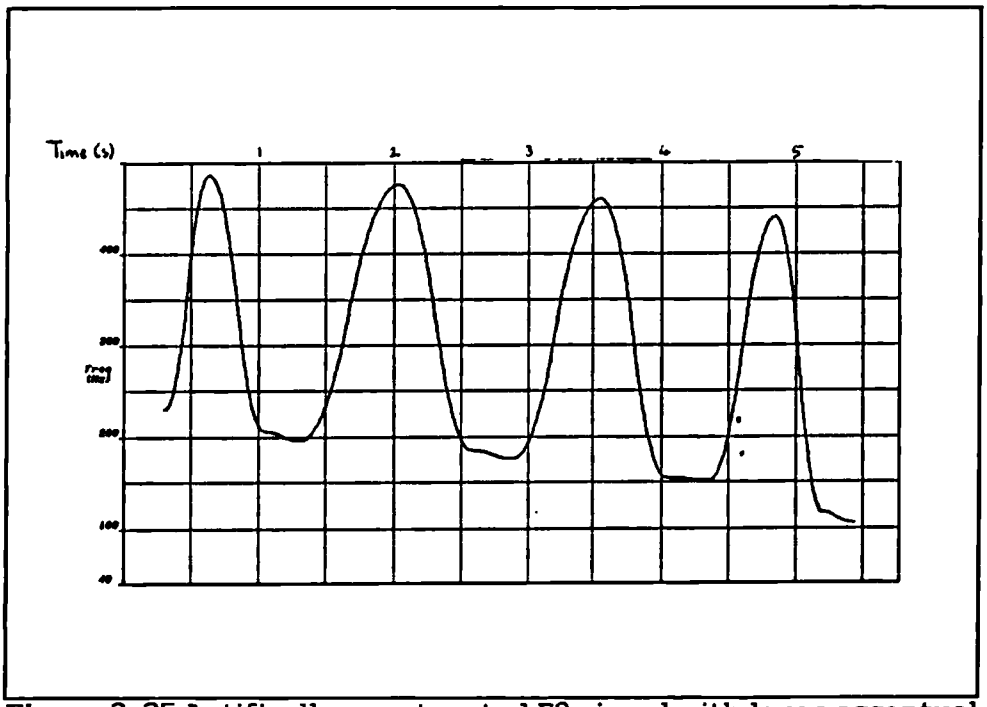


Figure 2.35 Artifically constructed F0 signal with large accentual excursions relative to that in Fig. 2.32.

excursions, and the effect of these can be seen to remain in the filtered signal, as a central hump, if the same window-size as in Fig. 2.34 is used (see Fig. 2.36). Clearly, a more robust method of retrieving the declining DC level is required.

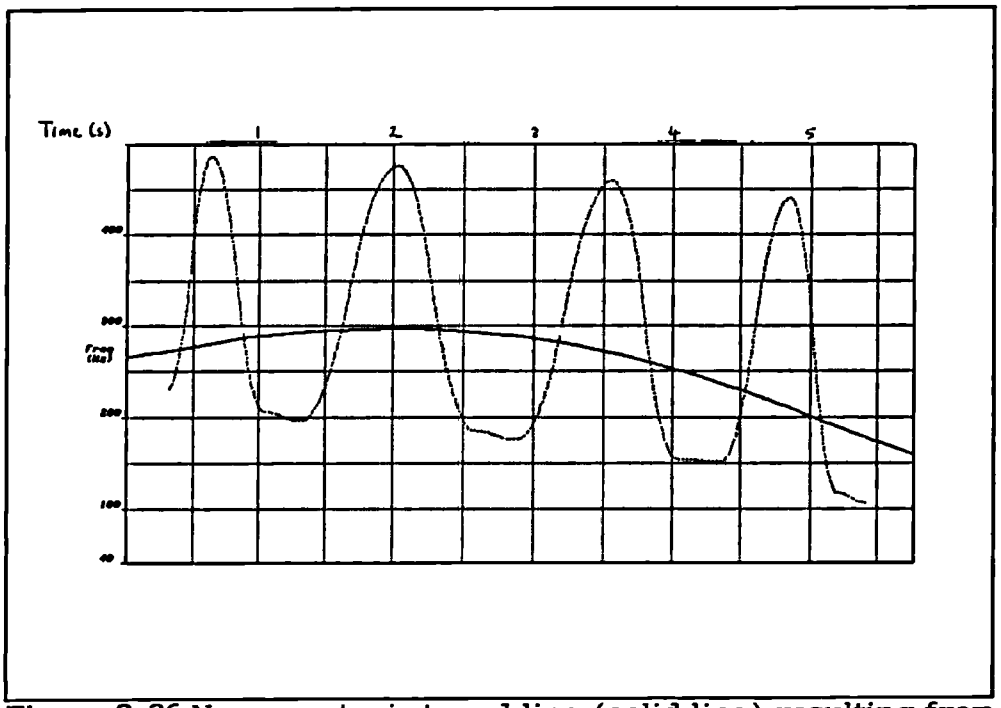


Figure 2.36 Non-monotonic trend line (solid line) resulting from single instance of convolution of 5700ms Hamming window with signal in Fig. 2.35.

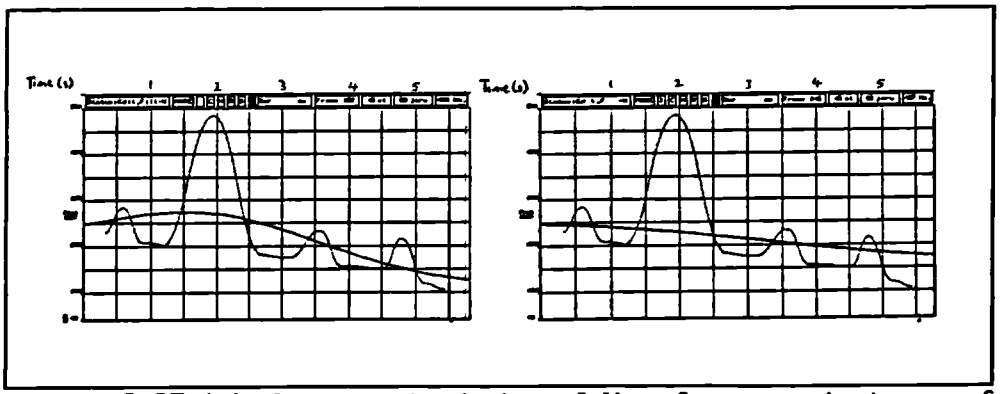


Figure 2.37 (a) Non-monotonic trend line from one instance of convolution of 5700ms Hamming window (signal has high 'amplitude' excursion on accent 2) (b) monotonic but shallow trend line after six such instances on same signal.

The problem with filtering the signal just as it is is that no criterion has been established for determining the effective narrowness of the filter bandwidth; there exists no yardstick by which it can be judged that the true angle of decline has been found for the declining DC level. Nor does it seem likely to be possible to guarantee that any such yardstick would not on some occasions result in the calculation of a declination line which is too shallow, because of the need to filter out high 'amplitude' local excursions (see Fig. 2.37).

An alternative approach is to determine that putatively 'true' angle of decline by first of all rotating the signal in the frequency-time (x-y) plane before filtering it with as wide a window[42] as is necessary to return a single DC level. The declination line can then be considered to be the DC line rotated back by the inverse of the original angle. All that is needed is some criterion by which the angle of rotation is considered optimal. Well, given that the DC level of the rotated signal can be considered to be a measure of central tendency, viz,. the mean value of that signal, an appropriate criterion is provided by mimimization of an associated such measure, the variance of the rotated signal sequence values. That is the criterion that has been adopted in the examples in Figs. 2.38 and 2.39. Fig. 2.38 shows the rotated versions of the three tone-unit contours which appeared in Fig. 2.4b, the original contours appearing as faint lines beneath. Fig. 2.39 shows the declination lines (dotted) retrieved by rotating back the computed DC line, with the declination lines calculated by linear regression (dashed) alongside for comparison.

[42] Typically, twice the duration of the signal

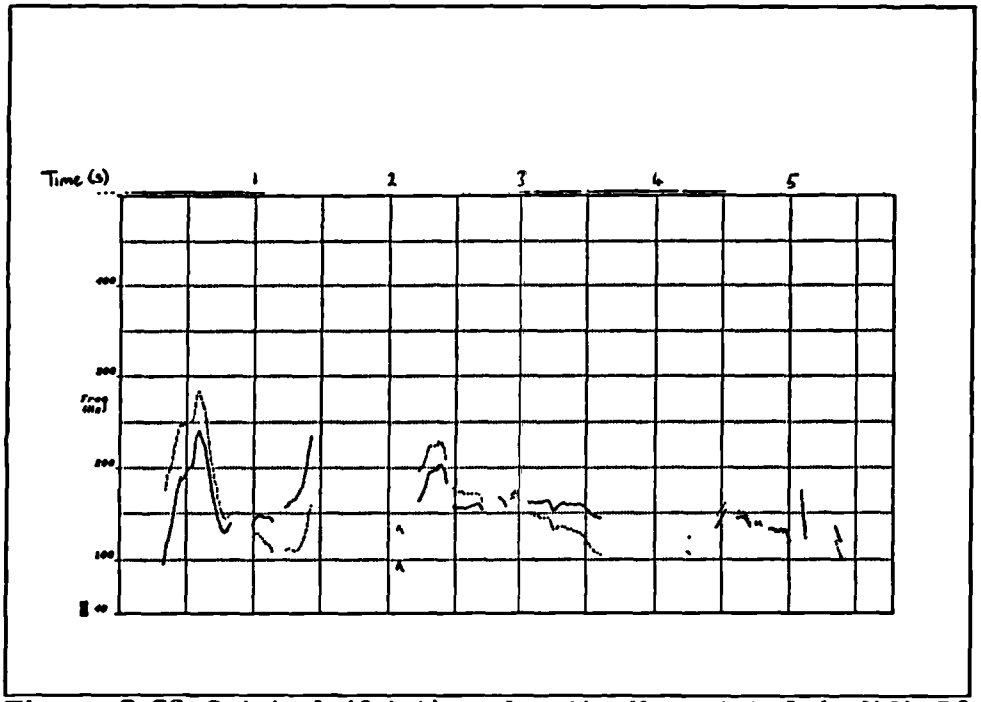


Figure 2.38 Original (faint) and optimally rotated (solid) F0 contours (cf. Fig.2.4b).

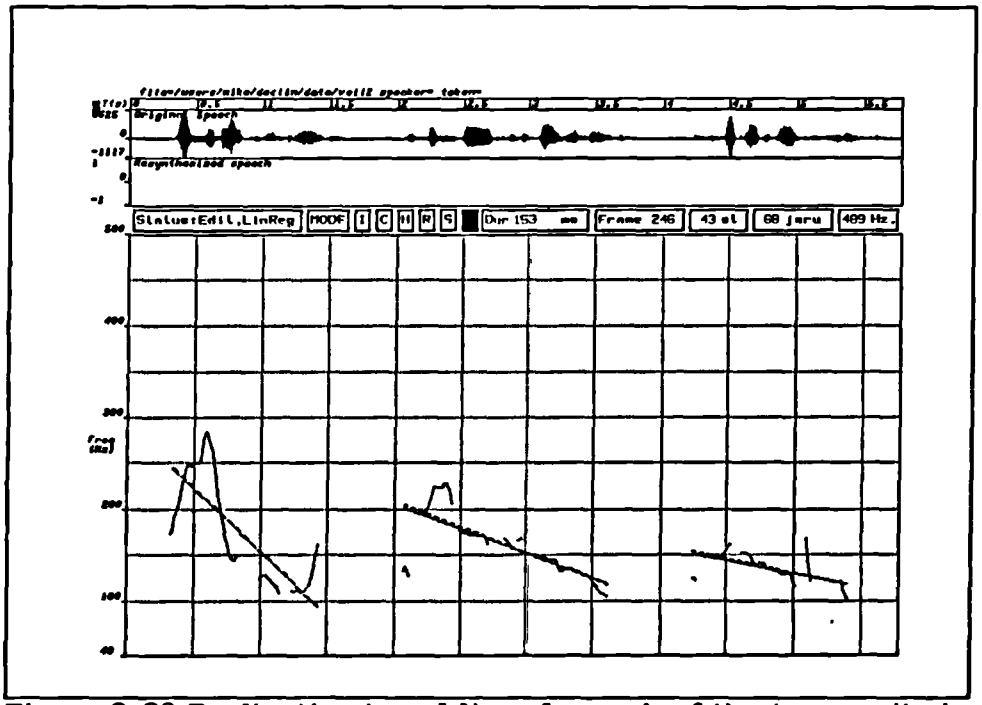


Figure 2.39 Declination trend lines for each of the tone-units in Fig. 2.4b. Dotted line: rotation method. Dashed line: linear regression

One or two comments about this procedure are in order. Firstly, the rotation is performed by pivoting the contour about its centre. For a contour of an

even number of samples, the pivot is between the two central samples, and for one of an odd number, it is the central sample. Secondly, because the contour is a time series, it is not possible to perform a standard geometric rotation in the x-y plane. Such an operation would entail non-linear distortion on the time-axis. The rotation that is performed is consequently simply effected by the following function:

$$f_i = f_i + \tan(x) * (i-p)$$

where f_i is the ith contour sample, x is the angle selected according to the criterion of minimal resulting contour variance, and p is the index of the pivotal sample. There necessarily results a distortion on the vertical axis which is not particularly evident at small angles of rotation, but at larger angles manifests itself as there appearing to be a greater rotation in stretches of the contour without local excursions compared with the local excursions in the contours themselves. For instance, in the contour of Fig. 2.40, where the angle of rotation is 0.5 rad., the local accents appear to be very similar in their orientation after rotation, having pretty much the same slopes as before, whereas slope of the unaccented stretches have more clearly shifted in value.

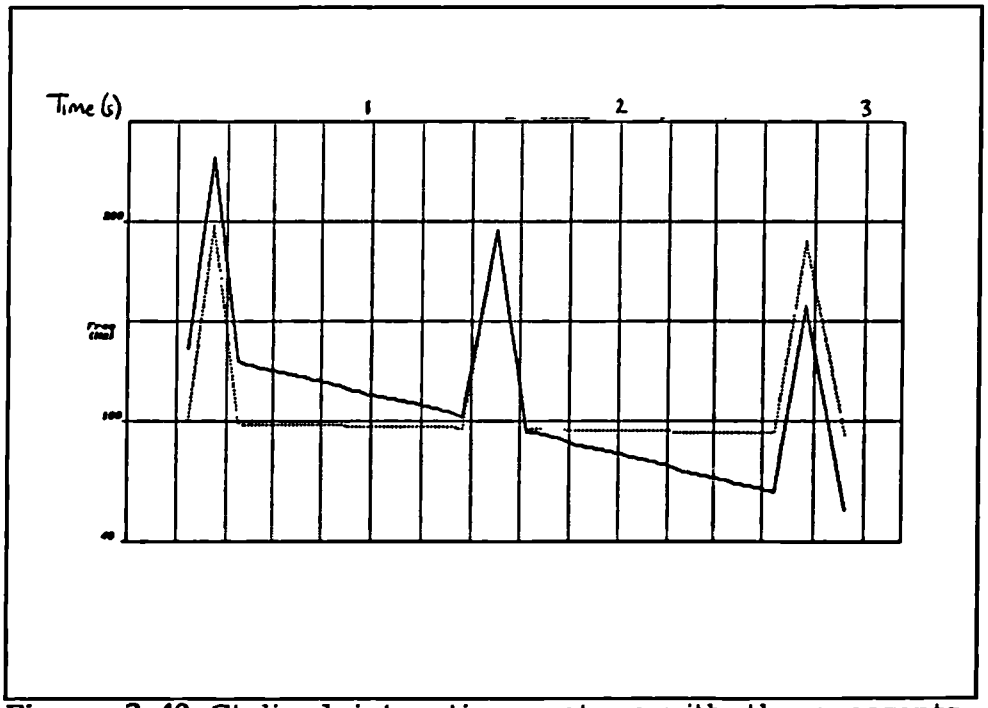


Figure 2.40 Stylised intonation contour with three accents. Original contour (dotted) and same contour rotated by an angle of 0.5 radians around the centre (solid).

In fact, the rotation method gives quite similar results to the method of linear regression using least squares (see Fig. 2.39). Indeed, it could be considered that it is merely a translation of the method of linear regression using a major-axis line, in which the squares of the perpendicular distances to the regression line are minimized. In addition, it must be recognised that the heuristic method used to determine the angle of rotation does invoke an algorithm which uses a criterial statistical measure of distribution, just as linear regression using least squares does. This digression has attempted to demonstrate the common ground that can exist between different methods of computing declination lines; in this particular case, that between linear regression methods and putative methods which might aim to retrieve a declining DC component in the F0 signal[43].

2.2.6 Fujisaki's analysis of declination

The rotation method demonstrated that the effects (in that particular case visual, but by extension, auditory) of high frequency components in a signal are less mutable than those of low frequency components when the signal undergoes time-domain distortion. For the types of intonation contour that were processed using the rotation method, there is a sense in which the accentual (high-frequency) components 'ride' on top of the 'intonation'/declination/inclination(low-frequency)component(s). Thisstate of affairs is exactly that obtaining in Fujisaki's model of Japanese intonation (e.g. Fujisaki and Hirose 1982, Fujisaki 1987), which is taken to be applicable in its basic form to other languages, including English.

Fujisaki proposes that an intonation contour comprises two components, an accentual component consisting of fast-moving positive F0 obtrusions and cooccurring with word accent in Japanese, and a phrase component, consisting of slower moving positive F0 obtrusions of generally smaller magnitude, and cooccurring with phrases, clauses and sentences. These are additive components, and superimposed on a non-declining baseline. Each phrase-component obtrusion is modelled as the impulse response of a second-order linear system with a relatively long rise time; each

---

[43]The analysis of a signal into DC and AC components might even be seen as the conceptual motivation for the division of the intonation contour into declination and accentual components.

accent-component obtrusion is modelled as the step-response of a separate second-order linear system with a relatively short rise time. The logarithm of the fundamental frequency at time t is given a value according to the following formula (taken from Fujisaki 1987, p.167) :

$$\ln F_0(t) = \ln F_{min} + \sum_{i=1}^{I} A_{pi} G_p(t-T_{0i}) + \sum_{j=1}^{J} A_{aj}\{G_a(t-T_{1j}) - G_a(t-T_{2j})\} \quad (1)$$

where
$$G_p(t) = \alpha^2 t.\exp(-\alpha t), \quad \text{for } t \geq 0$$
$$= 0 \quad \text{for } t < 0 \quad (2)$$

and
$$G_a(t) = 1-(1+\beta.t)\exp(-\beta t), \quad \text{for } t \geq 0$$
$$= 0 \quad \text{for } t<0 \quad (3)$$

F_{min} : asymptotic value of fundamental frequency in the absence of accent components
I : number of phrase commands
J : number of accent commands
A_{pi} : magnitude of the ith phrase component
A_{aj} : amplitude of the jth accent command
T_{0i} : timing of the ith phrase command
T_{1j} : onset of the jth accent command
T_{2j} : end of the jth accent command
α : natural angular frequency of the phrase control mechanism
β : natural angular frequency of the accent control mechanism

From these formulae, it can be seen that the intonation contour consists of the sum not only of phrase and accent components, but sequences of the same. This means that at any one time, the F0 contour can comprise the function (additive) of greater than or equal to one accent obtrusion, greater than or equal to one phrase obtrusion, and the F0 baseline. Of course, the difference in time of the successive terms T_{0i} and T_{1j} means that these components would be superimposed only at different stages of their respective 'lives'. In Figure 2.41 is given Fujisaki's example of the decomposition of an F0 contour into phrase and accent components, each sequence of which contains some overlap. The phrase and accent commands, which are the sequence of impulses and steps respectively which are supposed to correspond to individual commands to the laryngeal musculature, are derived from the phrase and accent components by deconvolution.

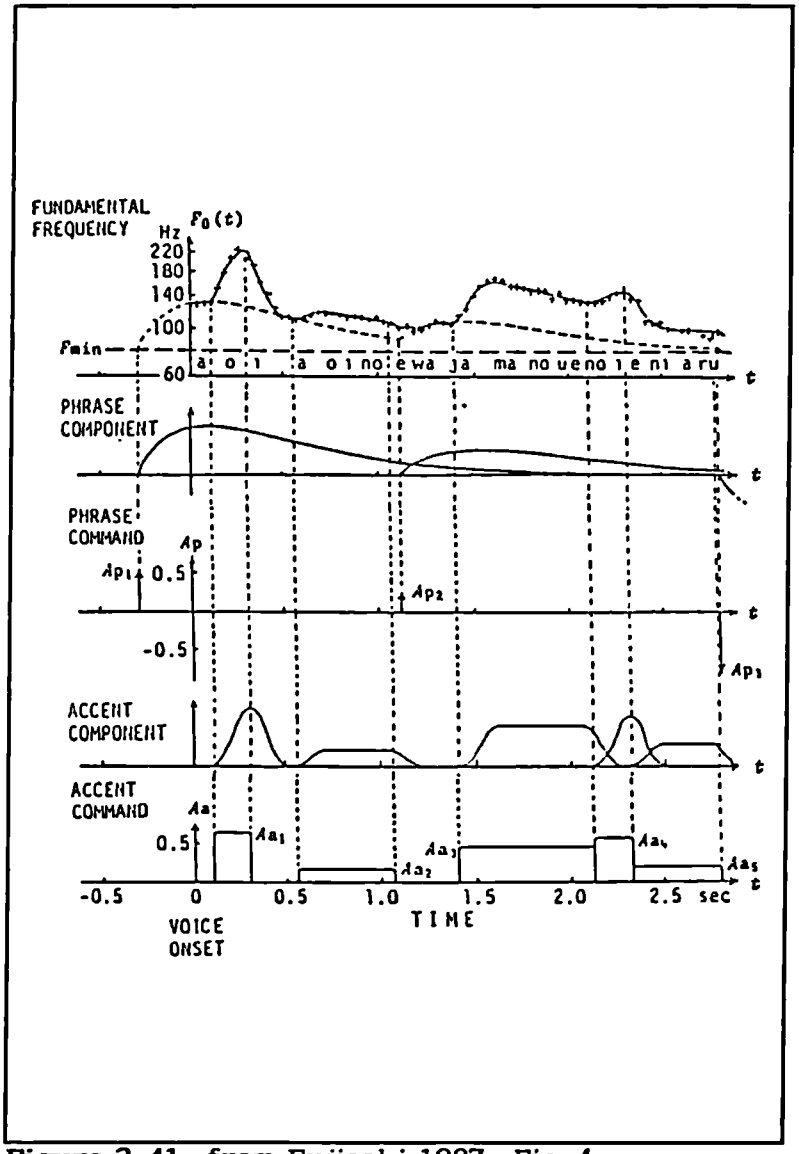


Figure 2.41, from Fujisaki 1987, Fig.4.

One of the advantages of this approach is that it does not require there to be a declining reference line throughout the whole of an intonation contour to account for declination. Declination, in so far as it is manifested as a slow decline in certain parts of the intonation contour, is attributable to the declining part of the phrasal component obtrusion. That declining part of the phrasal component also underlies stretches of F0 in which there is no overt declination, as a result of the additional accentual component. There are thus, if the baseline is ignored, parts of the intonation contour which are identical to declining parts of the phrase component, parts which reflect the decline of those parts directly in the decline superimposed on a flat-topped accentual component, parts which reflect that decline only indirectly and

marginally (and probably not perceptibly) by reducing the slope of accentual rises or falls; and then there are parts in which there can be no declination, because there is no decline at that point in the phrase component. Furthermore, the more intonation phrases there are in an intonation contour, the more phrase components, which will tend to overlap, reducing the amount of declination, or at least declination attributable to the phrase component[44]. Fujisaki's model thus makes interesting predictions about declination, not the least of which are (i) that declination occurs only in some sections of the intonation contour (as a function of the declining part of the phrase component), but that (ii) within those sections, declination is a global function, declining non-linearly in the F0 domain, and acting effectively as a declining baseline on which the more rapid accentual movements are superimposed.

2.2.7 Summary

This subsection has discussed mechanisms for computing declination lines, but in so doing has noted that there can be many interpretations of what declination actually is. The attribution of the whole of the downward trend in an F0 contour to a declination function has been considered, and contrasted with the approach in which much of that downward trend is partitioned into local functions of Initial Raising, Downstep and Final Lowering.

The former position is only acceptable if declination is considered to be a prosodic function which can be turned on or off; in the simplest sense, on for statements and off for questions with rising intonation (and rising main trend). This position is most consonant with the view that declination is not an automatic consequence of facts about the speech production mechanism during normal utterances and in particular of the tendency for subglottal pressure to decline during speech, but is under direct speaker control.

---

[44] The phenomenon which, according to the analysis in this chapter, would be interpreted as downstep, is also a separate phenomenon for Fujisaki. It would be manifested as a continuously variable reduction in the amplitude of consecutive accent commands or in the magnitude of consecutive phrase commands. This continuous variability is noted as a deficiency by Ladd (1990).

The latter position, that declination accounts for only a part of any downward trend in an utterance, and may account only for an attenutation of an upward trend, states that declination is a function of the frame of reference or coordinate system within which intonation contours are produced and perceived (cf. Ladd 1984). It is most consonant with the view that declination is an automatic consequence of facts about the speech production mechanism. There are many positions intermediate between and at variance with these two basic positions. At this stage, it is only possible to say that declination (qua a downward trend) is an obvious phenomenon in speech, but is not easily quantifiable, largely because its aetiology has not been clearly explored. It is one of the jobs of this thesis to develop an approach to investigating this aetiology.

2.3 THE DOMAIN OF DECLINATION

The discussion of methods of determining declination lines has led inexorably to a discussion of what precisely the phenomenon of declination is. Closely linked to both questions is that of the domain of declination. Thus far in the discussion, it has been assumed that the minimal domain of declination is the tone-unit or intonation phrase. In fact, it is the reset of putative declination lines at tone-unit boundaries which is often adduced as evidence for a particular view of the phenomenon of declination (Ladd 1990, 't Hart et al. 1990, Collier 1985, 1987). Declination functions are sometimes taken to have a domain which comprises two or more (minor) tone-units, as in Pierrehumbert 1980 and Pierrehumbert and Beckman 1988. A general rule of thumb appears to be that those treatments in which declination is a restricted declining frame of reference are likely to posit a domain of declination which is longer than those in which it is a more all-encompassing phenomenon with some necessary physical manifestation in the F0 contour. The former are likely to assume that the declination function is coterminous with a breath-group, (i.e. a prosodically well-formed utterance bounded by breaths), which is in keeping with the associated assumption that declination has a productive physiological aetiology. The latter are more likely to assume that there can be more than one declination domain within a breath-group. At the same time, higher order domains of declination are sometimes posited (cf. Thorsen,1985,1986).

The scope of the investigation in this thesis is such that the question of the domain of declination is not going to be addressed much beyond what is said in this subsection. The sentences used in Chapter 6 will comprise single intonation phrase[45] breath-groups uttered in isolation, so the issue of maximal domains of declination and declination resets will not arise, but will have to be addressed in later work.

One thing ought to be said about the minimal domain of declination, however. If an utterance has a number of accented syllables in it, and if it is assumed that there is still only one tone-group in the utterance, and if there is an identifiable 'declination effect' in the perception of the prominence on those accented syllables (that is, the greater the index of an accented syllable, the greater the prominence-F0 ratio), what happens if the utterance is curtailed before the speaker reaches the end of it - say, before the nuclear tone? Is there a global declination function identifiable over the unfinished utterance? The answer must be yes, if the declination effect has been seen to occur on prenuclear accented syllables. Yet if that is so, then such a declination function must be identifable over a curtailed utterance of arbitrary duration, and the question of the minimal domain of declination remains open, at least at the performance level.

The fact that this problem can be posed is an indication that more detailed work needs to be done to identify what phenomena, at a local level in an intonation contour, can give rise to the declination effect, and hence, what at a local level consitutes the presence or absence of a global declination function in both production and perception. What needs to be done, in short, is to identify the way in which a global declination function might be constructed from local phenomena of varying domains.

2.4 CONCLUSION

This chapter is concluded by a consideration of three main themes which have arisen out of it and which are prerequisites to satisfying the requirement presented at the end of section 2.3. One of the more problematic issues in

---

[45] An 'Intonation phrase' is taken here to be the same as a 'Major tone-unit' for O'Connor and Arnold (1973). It can be subdivided into 'Intermediate Phrases' (Pierrehumbert and Beckman (1988) or 'Minor tone-units' (O'Connor and Arnold 1973).

determining what declination is and how to measure it is the degree of abstractness that a putative declination function may have. That is, to what extent is it likely that a declination line or lines of reference can be established as a theoretical construct which could be used for scaling individual instances of F0 contours, without there being any physical manifestation of that reference line in the F0 contours, or transforms of them? The discussion in this chapter has shown that many approaches to modelling intonation contours involve such an abstract conception of declination. Some of these also account for much of the downtrend in an intonation contour by a process of downstep (see Chapter 3). A downstep function has clear physical exponents in the F0 contour, and thus could be argued to be less abstract than a putative declination function. However, for such a local function, the constraints on direct physical representation in the F0 signal are less severe than for a global declination function. Furthermore, as the discussion of Pierrehumbert's model of English intonation in the next chapter will demonstrate, the arguments for a downstep rule specifically of phonetic realisation have still to be made. So it is still conceivable that a rule of downstep might apply to more abstract categories in the intonological system of a language.

As for the other local downtrend processes, Initial Raising and Final Lowering, they are perhaps more clearly not abstract, since they seem to correspond (arguably, in the case of Initial Raising) to local physiological functions directly related to the frequency of vocal fold vibration. Thus, those processes involved in the downward trends in intonation contours are neither necessarily abstract nor, to adopt Cutler and Ladd's (1983) terminology, necessarily 'concrete'. To return specifically to the putative global declination line, amongst those treatments where it is posited, some treat it as an abstract declining reference line, and some as a concrete declining reference line. But the latter treatments (such as those of the Eindhoven School or Thorsen) must allow for a certain amount of abstraction, at least, during accented syllables; and indeed, there may be only a couple of points in the F0 contour which identify the declination line[46].

---

[46] It is often the case that the ends of final falls are used to identify the end of a declination line. Such 'concrete' approaches are thus usually the ones that conflate declination and Final Lowering, or deny the existence of the latter.

Where there is less abstractness in accounting for declination, there is more intonational componentiality involved in the analysis of F0 contours. That is, the F0 contour is seen as comprising the function of a number of contributing contours (all in the fundamental frequency domain), all of which have some physical manifestation in the actual contour produced and perceived. For this state of affairs to have some physiological veracity, one or both of the following ought to be the case: (i) there are different channels in the production of intonation contours which correspond to the different components in the F0 contour (for instance, Fujisaki (1987) proposes that contraction and relaxation of one part of the cricothyroid muscle (pars obliqua) is responsible for his - partially declining - phrase component, whereas that of the other (pars recta) is responsible for his accentual component); (ii) there are different channels in the perception of intonation contours which also correspond to the different components in the F0 contour. These states of affairs suggest that the global declination line could be a function of peripheral physiological or neurophysiological processes, and be produced and detected by them.

Both the more abstract and more concrete views of declination maintain the globality of the declination function over a particular domain. There is an alternative way of looking at things, and this is to suppose that at one level, declination is physically manifested in the F0 contour, but only in those stretches in which there is a gradually declining stretch of F0. This attribution of the locality of declination is the third and last of the main issues to be raised by the discussion in this chapter. According to this hypothesis, at the most peripheral level, there is only a single channel both for the production and perception of intonation, but there are separate detectors or generators for different F0 configurations in the time domain (e.g. accentual patterns, or declination patterns). That is to say, the F0 contour is separated into temporally abutting components rather than components superimposed in the frequency domain, by productive and perceptual processors.

This hypothesis is consistent with the hypothesis that there is a global declination function which is more abstract than requires physical manifestation in an F0 contour. It is possible that there are local declination generators and detectors which form the output and input respectively of

concretising and abstracting processes which, with the possible aid of additional cues from additional channels of information, are responsible for more abstract declination line constructs. It will be interesting to explore what such processes might be, but their existence is not required by the locality hypothesis. Local stretches of declining F0 may be produced, and the declination effect be shown primarily to be a local effect, without the intervention of higher levels of abstraction. The investigation of these hypotheses occurs later in the thesis. Before that, a more detailed investigation is made of the process which is suggested by Pierrehumbert to account for much of the downward variation in F0 contours, that of downstep.

CHAPTER 3
THE PHENOMENON OF DOWNSTEP

3.1. INTRODUCTION

The concept of downstep was first used in the description of African 'terraced-level' tone languages (e.g. Clements 1979). In those languages, a High tone following a sequence of High and Low tones is uttered at a markedly lower pitch than that of the High tone in the preceding sequence, and in some cases the intervening Low tone is elided or 'floats' (which is the standard case of 'downstep', according to Clement's 1979 terminology; with no elision, the phenomenon is referred to as downdrift).

This concept of downstep has been adopted by Pierrehumbert (1980), who derives a similar rule for scaling H tones following a HL sequence. The formal mechanisms she invokes differ in a number of ways from those used in accounts of African tone languages. In later work she has done in conjunction with Beckman on Japanese and English (Beckman and Pierrehumbert, 1986 and Pierrehumbert and Beckman, 1988), the link between the process in the African tone languages and that in non-tone languages, such as English or Japanese, is weakened. The formal context for downstep is extended, and the process renamed as 'catathesis', to distinguish it from the process in the tone languages. Here the term 'downstep' is retained, as it is more clearly descriptive and more widely used.

3.2 DOWNSTEP AS INTRODUCED IN PIERREHUMBERT'S 1980 THESIS

A study of the formal conditions for downstep in Pierrehumbert's model of intonation is useful in illustrating certain difficulties involved in identifying, in particular contours, the locations where downstep has taken place. These difficulties are worth it for Pierrehumbert, because the postulation of the process of downstep in English is pivotal in her 1980 account (as well as subsequent accounts) of the phonology of English intonation. In the first place, it is the downstep rule which allows her to specify that there are only two distinctive tone-levels, as opposed to three or four, in the intonation system of English[1]. The critical cases in making this decision are contours

---

[1] The use of static tones with interpolation to characterize English intonation contours continues the trend of intonation analysis due to Americans such as Pike (1945), Trager and Smith (1951) and

containing stepping heads (O'Connor and Arnold, 1973). The number of different steps in such heads is often quite large in list forms. One such, taken from Liberman and Pierrehumbert (1984), can be seen in Fig. 3.1. Here, there are five different accented syllables corresponding to five different pitch levels. Even if Initial Raising and Final Lowering are taken into consideration, there are three, and the possible number of additions to the list is of course infinite, allowing for the theoretical possibility of a large number of distinctive pitch levels being required in the intonation system of English on account of a rather small subset of utterable contours. The rule of downstep allows those different pitch levels to be generated from a sequence of elements of the dual system of pitch levels H(igh) and L(ow). In Pierrehumbert 1980, the rule applies to reduce the height of a High pitch level (H tone) by a constant factor of the scaled value of the preceding H tone. This mechanism ultimately produces the asymptotically declining contours of the sort seen in Fig.3.1 (if the final portion is discounted).

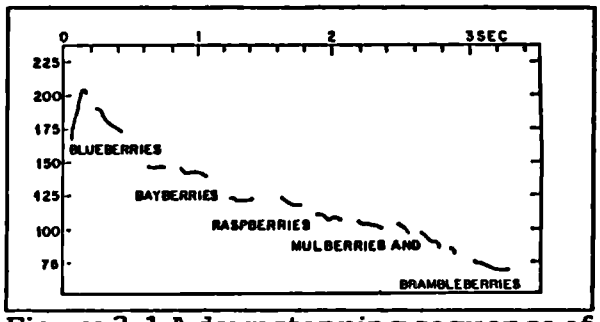


Figure 3.1 A downstepping sequence of accented syllables. Taken from Liberman and Pierrehumbert 1984, Fig.13, p.171.

In later treatments, the downstep rule also helps to secure the validity in her schema of the division of pitch accents (which are the 'molecules' of her tonal phonology) into monotonal and bitonal forms[2]. In the 1986 and 1988 papers, downstep serves to contextually distinguish monotonal and bitonal forms: all bitonal pitch accents trigger downstep in following material (in the case of Japanese, all immediately following material, even the second tone in the triggering bitonal pitch accent, and in the case of English, all material after the triggering bitonal pitch accent); on the other hand, no monotonal pitch accents trigger downstep.

Liberman(1975).

[2] This securement is also achieved by pointing out the requirement of descriptive adequacy in describing the detail of intonation contours - see Beckman and Pierrehumbert 1986, pp259-61.

The downstep rule also serves to underpin the concept of the phrase accent, adopted from Bruce's (1977) analysis of Swedish by Pierrehumbert into her 1980 thesis. In what Ladd (1978 and 1980) refers to as 'stylised' intonation contours, the apparently mid tones reached at the end of the most common, vocative chants, are analysed by her as downstepped H tones (which constitute phrase accents) followed by a rule of boundary tone 'upstep'[3].

The generative core of Pierrehumbert's model of intonation is in the finite-state grammar used to produce intonation contours, schematised in Fig. 3.2.

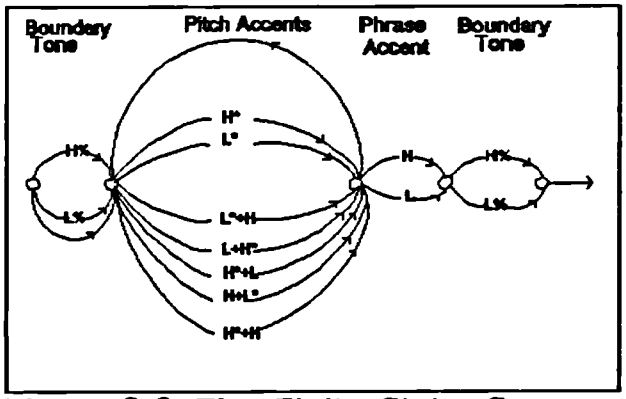


Figure 3.2 The Finite-State Grammar used to generate Intonation Contours in Pierrehumbert (1980)

The intonation contour produced in the generation of any intonation phrase comprises, in order, an initial boundary tone (if utterance initial) which is either H% or L%, an iteration of either bitonal (H*+H, H+L*, L+H*, H*+L or L*+H)[4] or monotonal (H* or L*) pitch accents, a phrase accent (H or L) and a terminal boundary tone (H% or L%).

From the discussion so far, it is apparent that the process of downstep is critically linked with the ontological status of all of the tonal forms except the boundary tone in Pierrehumbert's inventory. All these tones are required to describe accurately the form of the many different intonation contours which she includes in the appendix to her 1980 thesis.

Figures 3.3-11 (taken from Pierrehumbert's thesis) exemplify the relationship these tonal categories have with F0 contours. In Figure 3.3 (Pierehumbert's 4.29), a H(igh) initial boundary tone is shown, followed by a L(ow) monotonal pitch accent (L*), a H* monotonal pitch accent, a L(ow) phrase accent and a L(ow) final boundary tone. Reading the text with the

---

[3]See below for discussion of upstep.

[4] The theoretical possibilities H+H*, L+L* and L*+L are excluded as non-contrastive in Pierrehumbert 1980. In later work, H*+H is similarly excluded, so that there are no equipolar bitonal pitch accents.

tones aligined as in the figure, the contour can be heard to be a common iconic form. The rhythmically stressed syllables of the sentence are (at least) /ri:/ or "really" and /tru:/ of "true". It is to these syllables that the 'starred' tones of the monotonal pitch accents are attached. In fact, the interpretation to be placed upon the stars is that any tone with such a diacritic attaches to a stressed syllable in the process of the alignment of the 'tune' with the 'text'.

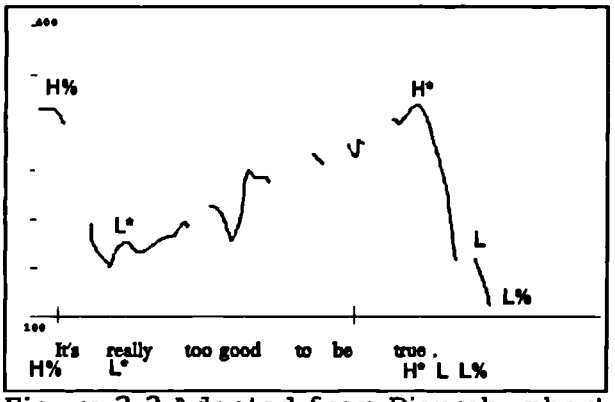


Figure 3.3 Adapted from Pierrehumbert 1980, Fig. 4.29, p.348

In Figure 3.4 (Pierrehumbert's 2.2), the independence of the tune and the text is shown by the alignment of the same sequence of tones with a different text.

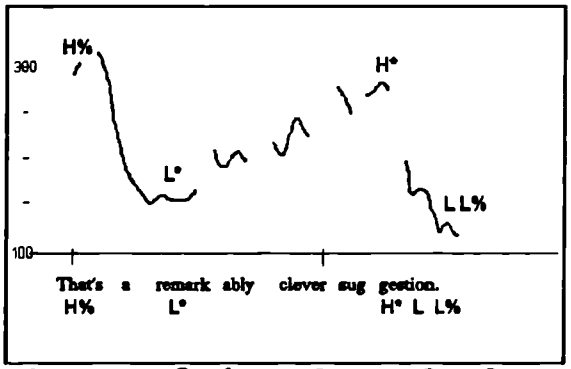


Figure 3.4 Adapted from Pierrehumbert 1980, Fig. 2.2, p. 276.

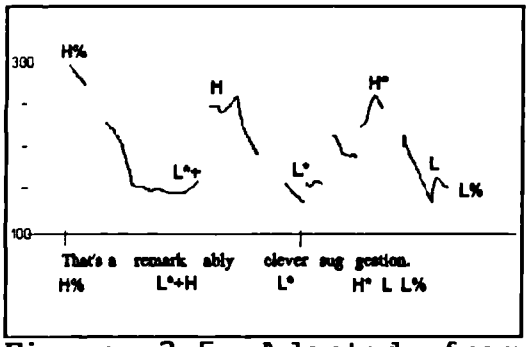


Figure 3.5 Adapted from Pierrehumbert 1980, Fig. 4.27, p.346.

In Figure 3.5 (Pierrehumbert's 4.27), the same text is aligned with a different sequence of tones, this time including a bitonal (L*+H) pitch accent. The starred tone of this pitch accent aligns with the stressed syllable, but it is the unstarred tone which has relatively high F0. In Figure 3.6 (Pierrehumbert's 4.4), a sequence of L+H* accents appears. These are differentiated from a string of H* accents by having bigger dips in between the peaks.

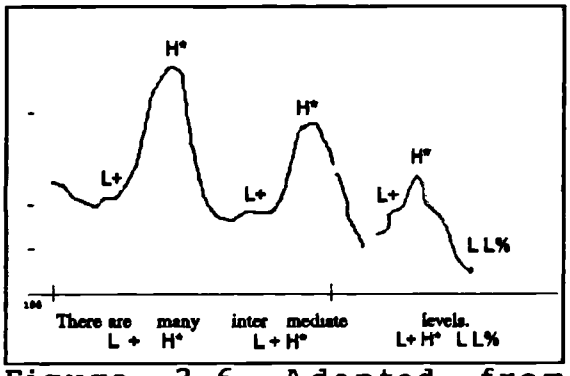


Figure 3.6 Adapted from Pierrehumbert 1980, Fig. 4.4, p.330

The downstepping sequence of Figure 3.7 (Pierrehumbert's 4.5) illustrates repetitive use of the H+L* pitch accent. The H*+L pitch accent is exemplified in Figure 3.8 (Pierrehumbert's 4.6).

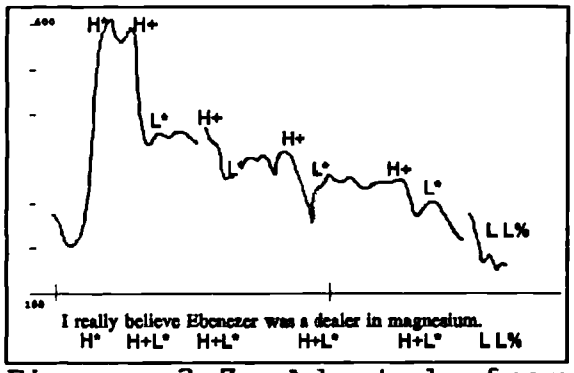


Figure 3.7 Adapted from Pierrehumbert 1980, Fig. 4.5, p.331

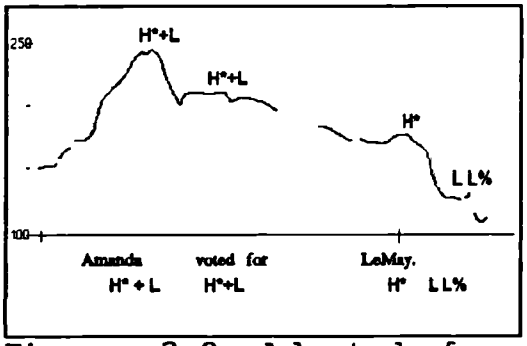


Figure 3.8 Adapted from Pierrehumbert 1980, Fig. 4.6, p. 332.

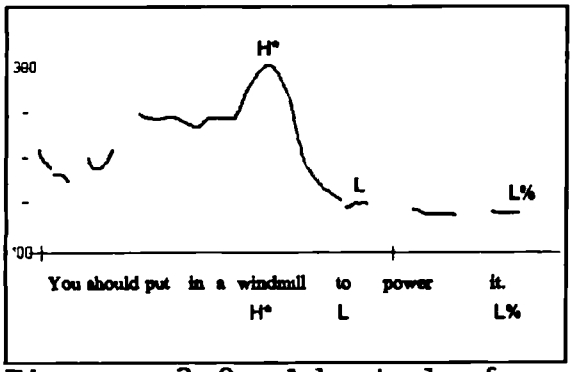


Figure 3.9 Adapted from Pierrehumbert 1980, Fig.2.32B, p.296

In all of the current examples, the phrase accent has been an L tone, and has seemed inseparable from the final boundary tone, which has always been L too. Figure 3.9 (Pierrehumbert's 2.32B) shows a more clearly demonstrated L phrase accent and L% boundary tone. In Figure 3.10 (Pierrehumbert's 1.13), the L phrase accent is seen in conjunction with a H% boundary tone.

In Figure 3.11 (Pierrehumbert's 1.8), H phrase accents are seen in conjunction with L% and H% boundary tones. The height of the L% boundary

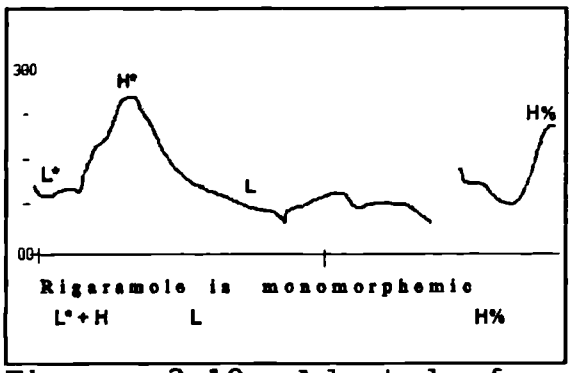


Figure 3.10 Adapted from Pierrehumbert 1980, Fig.1.13, p.267

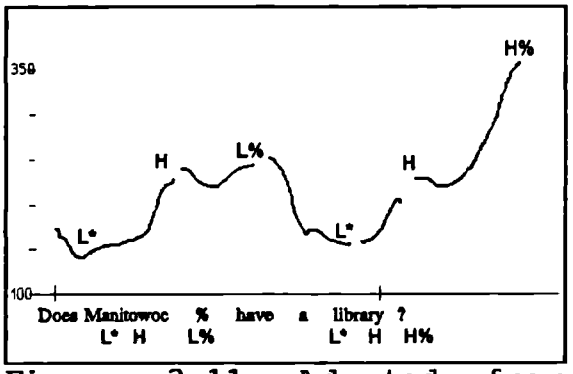


Figure 3.11 Adapted from Pierrehumbert 1980, Fig. 1.8, p.265

tone is different from that in earlier accents because of the H phrase accent, which triggers a rule of upstep (see below).

In all of these examples, Pierrehumbert considers it reasonable to align the tonal symbols with the F0 contour, because, for her, the F0 contour is the 'phonetic representation of intonation' (1980, p.13). That is, there is no intervening level of description between the phonological level (comprising the tonal symbol string) and the F0 contour; the phonological and phonetic rules listed below make quantitative predictions about the form of intonation contours which are sufficiently precise (even bearing in mind the effects of segmental coarticulation) that the F0 contour can be used to represent their output.

This view is reiterated in general form in Pierrehumbert and Beckman (1988): "What is special about phonetics is its use of quantitative information. Phonetic rules refer not only to linguistic categories and structures but also to properties of sounds and articulatory gestures. These must be represented using continuously variable numeric values of physical and psychophysical parameters" (p.1); and "[phonetic rules] differ from phonological rules in the representations they manipulate. They take as input phonological representations, but their output consists of quantitative functions, representing facts about articulations or sounds." (p.5).

The F0 contour is thus used as a repository for information about articulatory behaviour on the one hand and perceptual behaviour on the other; it is used as a unifying medium for the analysis of the process of communication.

The use of the F0 contour is certainly admissible for descriptive study of the intonational phenomena that speakers of utterances produce, and it is in this way that F0 contours are used in this thesis. In addition, although it seems odd that it should be further used as the formal output of a set of rules operating on categorical elements of intonational phonology, such use turns out to be consistent with the pursuit of concurrent physiological and psychological research[5]. Pierrehumbert's work is in the tradition of post-structuralist linguistic research pioneered by Chomsky, in which a clear distinction is made between competence models and performance models of language. It is taken that a model of the competence of a speaker of a language consists of no more than a model of the language; hence the concept of the 'grammar' of a language belonging to an individual. Evidence bearing on the particular form of a grammar in respect of certain syntactic considerations has always consisted of both observed and observable utterances. Chomsky himself tended to rely on observable utterances, as underwritten by a speaker's intuitions about the acceptability of their form. But this data, what Chomsky called 'empirical' data investing an empirical enterprise, is used to decide on the form of what boils down to a purely computational description of the language.

Pierrehumbert and others within the discipline of experimental phonology hope also to clarify the form of the phonetic part of this computational description of the language; they cannot use as their empirical data pitch contours as perceived by an introspecting informant, because most introspecting informants filter out of their perceptions or imaginations of intonation contours certain important phonetic facts bearing on physiological and psychophysical aspects of intonation (such as declination). It is for this reason that F0 contours are used as the raw data, and why it is considered acceptable to tag the contours with categorical symbols, just as the orthographic representation of a sentence is given a labelled syntactic bracketing. However, it has always to be remembered that Pierrehumbert's phonological and phonetic rules remain part of the description of the competence of a speaker of a particular language (and the important point that phonetic rules are language-specific, too, needs to be underlined). They can support and be supported by descriptions of the intonational

[5] What remains a difficult issue is the way in which results of phonological research can corroborate those of the latter fields, and vice versa.

performance of a speaker of a language, whilst acknowledging the above noted problems in that enterprise; it is a partial such description which is attempted in this thesis[6].

Returning to the contours just exemplified; they are generated by phonological rules (divided into main and tonal adjustment rules, the latter comprising the downstep and upstep rules) which determine the phonetic value of the individual tones of the pitch accents, the phrase accent and the boundary tones which comprise the pivot points in an intonation contour; these are supplemented by phonetic tone spreading and interpolation rules which produce the continuous F0 contour. The following constraints apply to the rules:

1. The rules apply left to right on the bitonal and monotonal entities in the intonation string.
2. The main phonological rules can make reference (a) to the phonological structure, and (b) to the prominence value, both of an entity for which a phonetic value is being computed and of the entity preceding it. The other rules can make no reference to the prominence value of either entity.
3. The rules can refer to only those two entities; that is, the rules are local rules, and exhibit no non-local dependencies.

---

[6] The use of F0 contours as the raw data of phonetic and phonological research into intonational competence can still be objected to on the grounds that the F0 contour is affected by too many accidental aerodynamic and acoustic phenomena, and indeed analytic artefacts. Their use as the raw data of such research into intonational performance can be objected to on the grounds that the F0 contour is an acoustic entity, and the acoustic domain ought not to have a favoured status in the study of human communication. This is because the signal which is used as a reference in both speech production and perception is a neurological signal, not an acoustic signal. We have no way of linking up speech production and perception via the acoustic signal wihout some assumptions about the speech production and perception mechanisms. We can link up speech production and perception using an internal neurological signal without making any assumptions about the acoustic signal, though we can use the acoustic signal as evidence for a modelled neurological signal; any model of speech communication using a modelled neurological signal has also to be consistent with the acoustic facts. This approach is attempted in this thesis. What can be considered heuristic approximation to averaged neurophysiological signals appears in work by the Eindhoven School and Hirst (e.g. Hirst 1983), in which contours are stylised.

4. The phonetic value of each tone (be it a tone in a bitonal or monotonal pitch accent, or in a phrase accent or boundary tone) and intervening stretches is computed once in strict sequence.
5. The cycle of rule application over the domain of a single tone incorporates all rule types, although the spreading and interpolation rules (which are mutually exclusive and contextually determined) apply with a single tone delay. Thus, once the rules to determine the phonetic value of a tone are begun, the interpolation and spreading rules will have applied to determine the continuous contour up to the point in time of that tone's predecessor.

The next constraint is one which appears to apply, but there will be seen to be problems with it:
6. If the context for a particular tone in the sequence is not satisfied in the Structural Description of any of the rules, then no phonetic value is computed for that tone.

Following is the set of rules, set out according to category. The phonetic value computed for each tone is computed in 'baseline units', which quantity is a linear transform of F0 depending on a globally declining baseline (see Chapter 2, Section 2.2.4.2). In Pierrehumbert's, account, modelling of the phenomenon of declination is restricted to this global frame of reference, grosser aspects being accounted for by downstep. A considered analysis of downstep is necessary, separate from one of declination proper in her treatment, to see how much of the descriptive load of the downward trend in intonation can be borne by it.

3.2.1 Pierrehumbert's 1980 rule set

(The omissions in numbering reflect the fact that certain rules have superseded more preliminary versions. Rule 2 has superseded rule 1, and Rule 10 has superseded Rule 3)[7].

---

[7] The superseded rules, 1 and 3, are as follows:

Rule 1.

$$\text{In } H^*_i \; H^*_{i+1}: \quad /H^*_{i+1}/ = /H^*_i/ \frac{\text{Prominence } (H^*_{i+1})}{\text{Prominence } (H^*_i)}$$

This rule Pierrehumbert claimed to be too specific in its context. In particular, the same relative scaling of the phonetic value of an H tone was found to be applicable to unstarred tones in bitonal pitch accents and phrase

A. Main phonological rules

(Here, H refers to a H(igh) tone, L to a L(ow) tone, and T indiscriminately to a H or L tone. Slash brackets ('/') mark reference to the phonetic value of the enclosed tone; round brackets indicate optionality, as normal. The Structural Description is the context; in the rules, what appears before the colon).

For H tones:

Rule 2. In H_i(+T) (T+)H_j : $/H_j/ = /H_i/$ Prominence (H_j)
Prominence (H_i)

(In this rule, only the immediate context is given, we must assume from Pierrehumbert's text. Thus, the sequences

$T+H_i\ H_j+T$,

$T+H_i\ T+H_j$,

$T+H_i\ H_j$,

$H_i+T\ H_j+T$

and $H_i\ H_j+T$

also fit the Structural Description.)

Added convention: If H_j is a phrase accent, then Prominence (H_j) = Prominence (H_i). As a result, in that case, Rule 2 predicts that $/H_j/ = /H_i/$.

---

accents; for completeness, it was also extended to apply to H% boundary tones, hence the general form of its replacement, Rule 2, but Pierrehumbert was unsure of the general validity of that extension (1980, p.145).

Rule 3.
In H+L: $/L/ = k\ /H/$ $0<k<1$

This rule Pierrehumbert found she had to turn into the more specific rule 10, because of a difficulty arising from one particular context triggering the downstep rule, viz. H*+L H, in which the L of the bitonal pitch accent has no phonetic realisation, but which, according to Rule 3, would get an L value computed for it which would require interpolation to, only for the interpolation to be nullified by a later interpolation rule which would have to bypass the unwanted L pivot. This rewrite of the contour is not allowed by constraint 4 above. Rule 10 excludes the H*+L H context, so that no rule exists for specifying the phonetic value of L in this context. According to constraint 6 above, no value is thus generated, and the L tone is deleted prior to interpolation.

For L tones:

Rule 10. In H+L* : /L*/ = k /H/ where 0<k<1

(k is a constant, operative at least for the whole contour. The value k is the same in Rules 10 and 8).

Rule 4. Where L is part of a bitonal pitch accent ,

In H (+T) L+ : /L/ = n /H/ $\frac{\text{Prominence (H)}}{\text{Prominence (L)}}$ where 0<n<k

Rule 5. Where L is a phrase accent ,

In H (+T) L : /L/ = p /H/ where 0<p<k

Rule 6. /L%/ = 0

(This means that a L(ow) boundary tone is on the baseline).

Rule 7. Where L_{i+1} is a pitch accent tone[8] or a phrase accent,

In L^*_i L_{i+1} : $/L_{i+1}/$ = $/L^*_i/$ $\frac{\text{Prominence } (L^*_i)}{\text{Prominence } (L_{i+1})}$

B. Tonal Readjustment Rules

Rule 8. (The Downstep Rule)

In H+L H_i and H $L+H_i$: $/H_i/ = k/H_i/$

Rule 9. (The Upstep Rule)

Where H is the phrase accent (and thus T the boundary tone) :

In H T: /T/ = /H/ + /T/

C. Tonal Spreading Rules (ordered before Interpolation Rules)

Where T is either the unstarred tone in a bitonal pitch accent or a phrase accent:

---

[8] Pierrehumbert does not use the qualifier 'tone' in her gloss of the context for this rule, but it seems reasonable to assume it was intended, since otherwise there would be no context specified in the set of rules given for the scaling of leading L tones of L+H pitch accents following L* pitch accents.

Rule S1. (the Rightward Spreading Rule)
T_i spreads towards T_{i+1} if $/T_{i+1}/ >= /T_i/$

Rule S2. (the Leftward Spreading Rule)
T_i spreads towards T_{i-1} if $/T_{i-1}/ = T_i$

(Rule S2 is a bit of an oddity as it postulates a leftward moving process. However, it does not violate the left to right implementation constraint, which is imposed on tonal entities rather than arbitrary segments of F0 contour. Leftward spread between tones would constitute a case of phonetic anticipation, and in a computational implementation of the rules could be implemented by a left-to-right sustention of values, since the spreading and interpolation rules lag the tonal evaluation rules by one cycle. Pierrehumbert is uncertain whether Rule S2 really exists in English).

D. Interpolation Rules
In contexts other than those specified in section C (spreading Rules) :

Rule I1.

In H_i H_j : Perform a sagging interpolation between H_i and H_j.

(In Pierrehumbert 1981, this is performed by a two-piece polynomial interpolation function).

Rule I2.

In T_i L_j and L_i T_j : Perform a linear interpolation between T_i and L_j or L_i and T_j.

Some examples of the implementation of these rules follow.

3.2.2 Examples and discussion of rules

In all these examples, the computation of targets (and possibly interpolating stretches, as appears to be the case in Pierrehumbert's Fig. 4.15) is done in baseline units. As the baseline is typically declining, this means that the contour appearing in (a) the first half of the figure would be computed as a downward-tilted version of the straight-line version of the contour in (b)

the second half of the figure[9] (along with adjustments made by rules accounting for segmental coarticulation effects). It should be noted, too, that all tones in the tonal string have prominence values associated with them as the result of assignments made on the basis of metrical strength and pragmatic emphasis (see discussion below). The values that are referred to in the course of the derivation of the values in baseline units which map into F0 values are thus always what Pierrehumbert refers to as 'phonetic values' rather than prominence values. However, it should be noted that in the absence of the application of any rule, the phonetic value is taken directly as the prominence value (for H tones) and the reciprocal of the prominence value (for L tones)[10]. It seems that this is what Pierrehumbert intended, although the current author has not been able to find any explicit statement to that effect in her thesis.

The constants in all these examples are:

k=0.6

n=0.2

p=0.1

Rule Implementation Example 1(Fig. 3.12)

In this example, four of the main phonological rules are exercised, along with one readjustment rule, one spreading rule and both interpolation rules. Firstly, the phonetic value of 1.5 is assigned directly from the prominence value to the initial H% boundary tone, and similarly 2.7 to the first H* tone. According to Pierrehumbert, the assignment to the initial pitch accent in the utterance is a free choice, and reflects the choice of pitch range for the whole utterance; it must be assumed that the assignment to the initial boundary tone is similarly free. The prominence value of 1.5 for the initial H% boundary tone is the same as its phonetic value. As noted above, this appears to be a convention, in the absence of current or previous modifying rules.

---

[9]The figures in (b) have been generated by the author following analysis.

[10]There is an apparent inconsistency here with constraint 6, but the author hasn't been able to resolve this.

Next, the sagging interpolation indicated by Rule I1 is applied between the initial boundary H% tone and the first tone in the H*+H pitch accent. Then, the phonetic value of the second tone in that pitch accent is computed. According to a convention, the prominence of an unstarred tone in a bitonal pitch accent is identical to that of the starred tone. Rule 2 thus applies again, trivially, to that second tone, to yield the same phonetic value of 2.7.

Next, the leading H tone in the second bitonal pitch accent is given a value of 2.7 by Rule 2, because it too has a prominence of 2.7. Then, spreading rule S1 operates to sustain the value of 2.7 between the two floating H tones. Rule 10 operates to compute the value of the associated L* tone in the H+L* pitch accent as 2.7*0.6 = 1.62, and linear interpolation according to Rule I2 takes place between the H and L* tones.

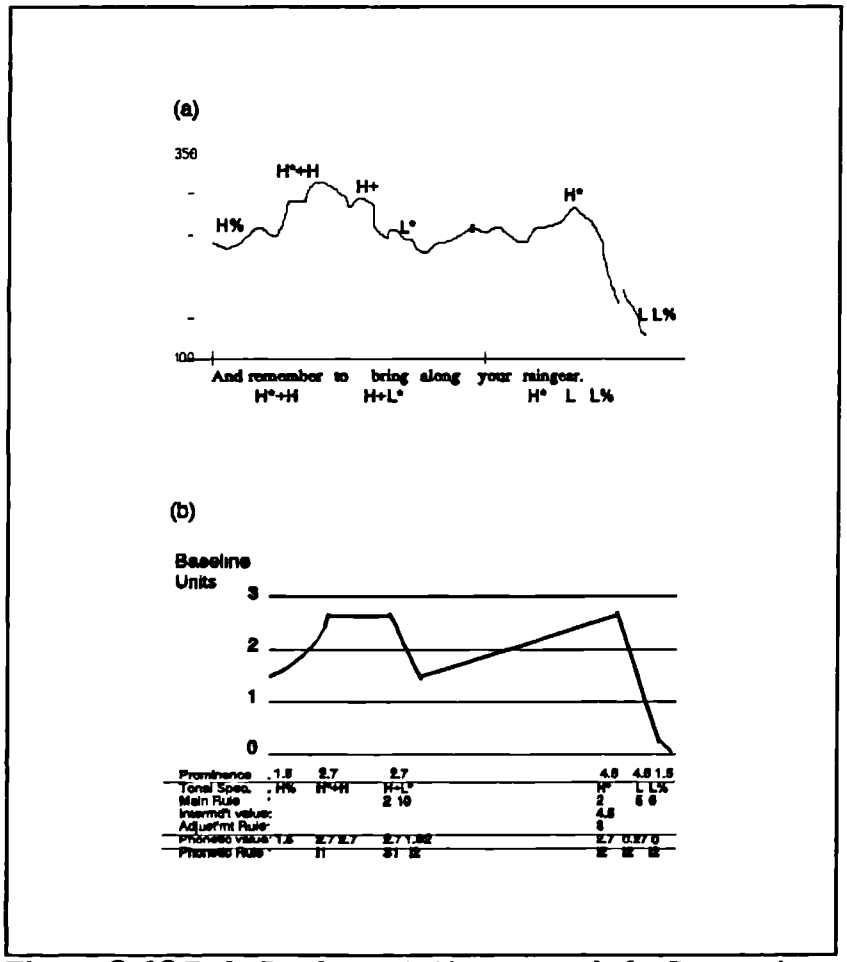


Figure 3.12 Rule Implementation example 1. Comparison of natural with rule-generated contour (a) Adapted from Pierrehumbert 1980, Fig. 1.15, p.268; (b) Stylised contour generated by rules, with synopsis of rules applied in derivation.

The final, monotonal H* pitch accent then has its value computed in two stages. Firstly, Rule 2 operates, giving the value 2.7*4.5/2.7 = 4.5. Then the readjustment rule of downstep operates to yield a value of 4.5 * 0.6 = 2.7. Thus, although downstep has operated on this H tone, it has the same phonetic value as its predecessor, because it has an appropriate amount of greater prominence. Interpolation by Rule I2 completes this step.

The L phrase accent has a value of 2.7*0.1 = 0.27 computed by Rule 5, and it is interpolated to by Rule I2. The L% boundary tone is assigned a value of 0 by Rule 6, and is also interpolated to by Rule I2.

Rule Implementation Example 2 (Fig. 3.13)

In the next example, a further one of the main phonological rules is exercised. The initial L% boundary tone is given a value of 1 directly as the reciprocal of the prominence value, and the L* tone of the following L*+H accent given a free value of the reciprocal of its prominence, that is 1/3 = 0.33. The H tone of that accent

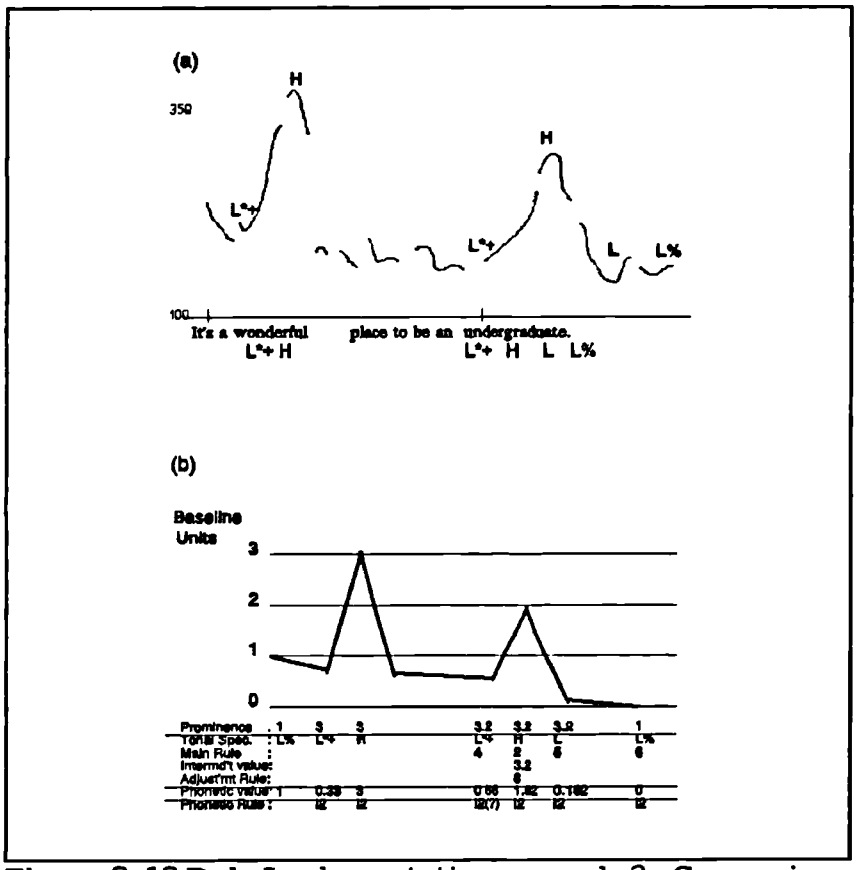


Figure 3.13 Rule Implementation example 2. Comparison of natural with rule generated contour: (a) Adapted from Pierrehumbert 1980, Fig. 4.34, p.351; (b) Stylised contour generated by rules, with synopsis of rules applied in derivation.

is given a free initial value of 3 directly from the prominence value[11]. The L* tone is given the value of 0.2*3*3/3.2=0.56 by Rule 4. The following H tone is given a value of (3.2/3*3)*0.6=1.92 by Rules 2 and 8. The L phrase accent gets a value of 1.92*0.1=0.192 by Rule 5, and Rule 6 assigns a value of 0 to the L% boundary tone. All assignment cycles bar the first are completed by interpolation by Rule I2[12].

---

[11] Although Pierrehumbert says that it is the initial assignment to a pitch accent which is free, there could be a case for arguing that the initial boundary tone could be given the sole responsibility of free assignment. Then the initial pitch accent in the phrase could have values computed by Rule 2 (for H*) with possible modification by Rule 8, or Rule 4 (for L*).

[12] There appears to be an error in Pierrehumbert's analysis of this contour. Interpolation between any two tones one of which is an L tone is one-piece linear, yet there is a dip after the first L*+H accent which is unexplained.

Rule Implementation Example 3 (Fig.3.14)

In this example, the remaining main phonological rule is exercised, and the other readjustment rule. The initial L% boundary tone, which is taken by the current author as a necessary starting point, is given a value of 0.83, the reciprocal of its prominence of 1.2, and the following L* tone is similarly given a value of 0.8, (the reciprocal of its prominence of 1.25). Rule 7 applies to the following L* tone, yielding a value of 0.8*1.25/1.5 = 0.67. The H phrase accent is given a value of 1.8 direct from its prominence, and the H% boundary tone gets an intermediate value of 1.5*1.5/1.5=1.5 after Rule 2 has applied, and a final value of 1.5+1.5=3 after the upstep rule, Rule 9, has applied.

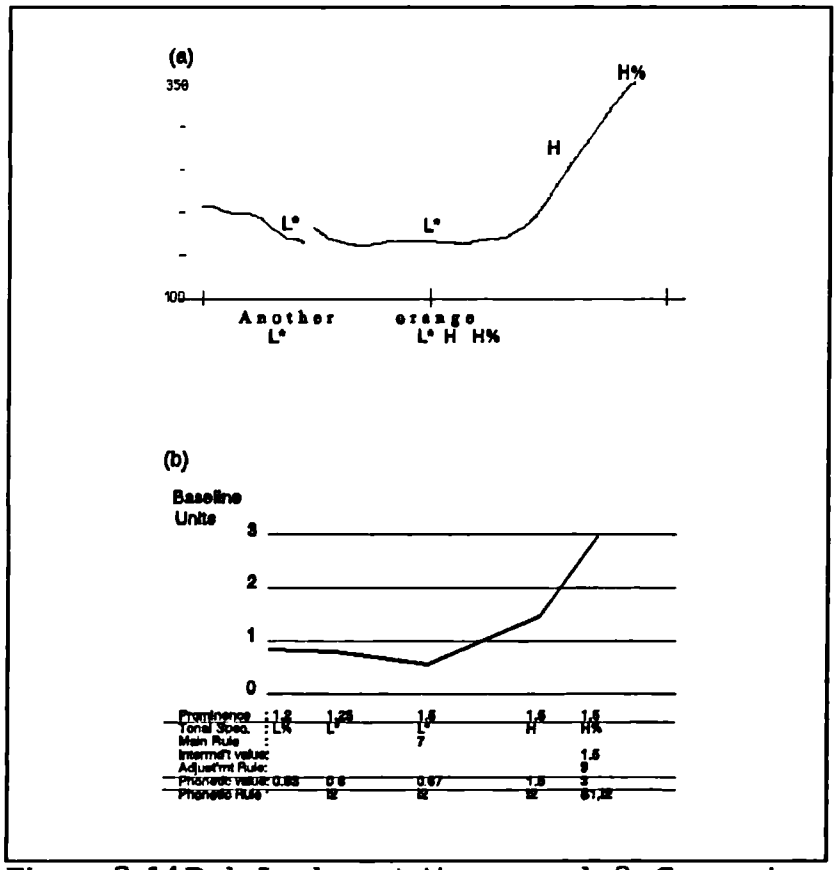


Figure 3.14 Rule Implementation example 3. Comparison of natural with rule-generated contour: (a) Adapted from Pierrehumbert 1980, Fig. 1.3B, p.259; (b) Stylised contour generated by rules, with synopsis of rules applied in derivation.

Rule Implementation Example 4 (Fig. 3.15)

This example shows Rule application to a sequence of H* monotonal pitch accents, in which the nonmonotonic 'sagging' interpolation rule is exercised twice. This example also demonstrates the existence of a downward shift in F0 between two pitch accents (the first and the second) in which downstep plays no part, and which is wholly attributable to the difference in underlying prominence of the two H* tones.

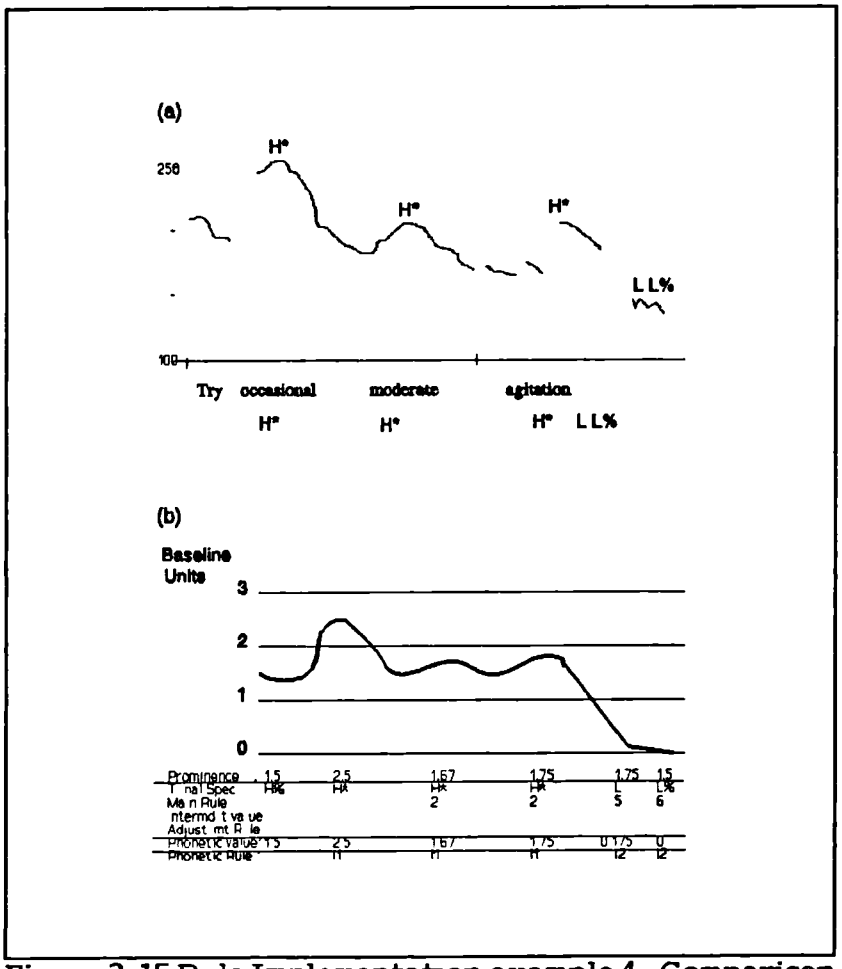


Figure 3.15 Rule Implementation example 4. Comparison of natural with rule-generated contour: (a) Adapted from Pierrehumbert 1980, Fig. 2.11, p.281; (b) Stylised contour generated by rules, with synopsis of rules applied in derivation.

Rule Implementation Example 5 (Fig.3.16)

This is the first of four examples demonstrating the downstep process. An initial H% boundary tone is given a value of 0.33, the reciprocal of its prominence value (3). It is followed by a monotonal pitch accent whose H*

tone is given the free value of 3 directly from its prominence. Next comes a H+L* bitonal pitch accent;

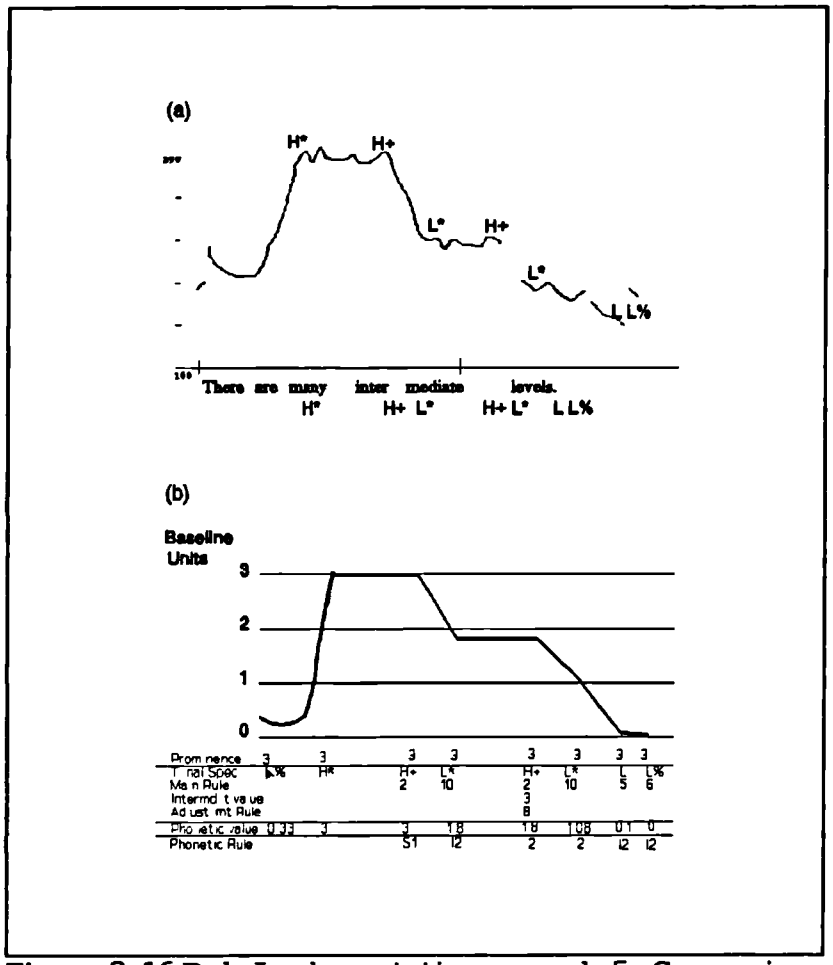


Figure 3.16 Rule Implementation example 5. Comparison of natural contour with contour generated by rule: (a) Adapted from Pierrehumbert 1980, Fig. 4.1, p.329; (b) Stylised contour generated by rules, with synopsis of rules applied in derivation.

the H tone is given the same value of 3 by Rule 2, because both it and the preceding tone have prominence values of 3. Note that here is a context in which the leftward spreading rule S2 has to apply, to account for the fact that there is no sag between the preceding H* tone and this H leading tone of the H+L* pitch accent. It is necessary, because the possibility that Pierrehumbert suggests (1980 p.233) that the position of the H leading tone in such cases could be further left, the lack of sag being accounted for by it spreading rightward can't be allowed here, since there can be no rightward spreading toward the L* tone by spreading rule S1 because the L*

tone ends up with a lower value than the leading H tone. The L* tone is given a value of 1.8 (=3*0.6) by Rule 10.

Another H+L* sequence follows. The H gets a value of 1.8, the same as the preceding L* tone, because, again, it has the same prominence as the preceding H, and because it is adjusted by the same factor in Rule 8 as appears in Rule 10. The L* gets a value of 1.08 (=1.8*1.6) by Rule 10. Finally, the phrase accent is given a value of 0.108 (=1.08*0.1) by Rule 5, and the boundary tone a value of 0 by Rule 6.

Rule Implementation Example 6 (Fig.3.17)

This is the second downstep example, which involves a sequence of H*+L tones, in which the L tone is not realised in the contour. The initial H% boundary tone is given an initial value of 1 (note that it could have been interpreted as an L% boundary tone - with a prominence value of 1, both L% and H% boundary tones come out with a phonetic value of 1. There follows a sequence of three H*+L bitonal pitch accents, in which the H* tones are all given values in the same way as in Example 5. Note that because no value is assigned to the L tones, they are not treated as targets for interpolation, and so interpolation rule I1 operates between the resulting pairs of H tones.

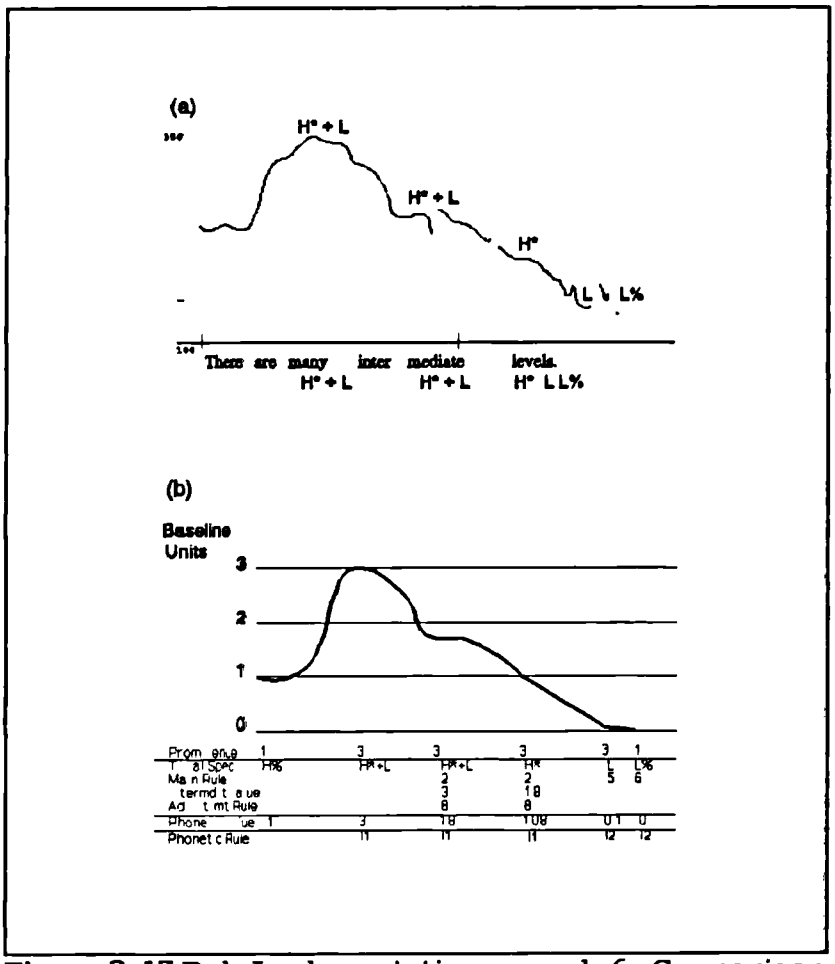


Figure 3.17 Rule Implementation example 6. Comparison of natural contour with contour generated by rule: (a) Adapted from Pierrehumbert 1980, Fig. 4.2, p.329; (b) Stylised contour generated by rules, with synopsis of rules applied in derivation.

Rule Implementation Example 7 (Fig.3.18)

The third downstep example involves a sequence of L*+H bitonal pitch accents, for which the L* tones other than the first (which is given a free assignment of 0.33, the reciprocal of its prominence) are assigned values according to Rule 4, and for which the H tones are assigned values in exactly the same way as in the previous examples.

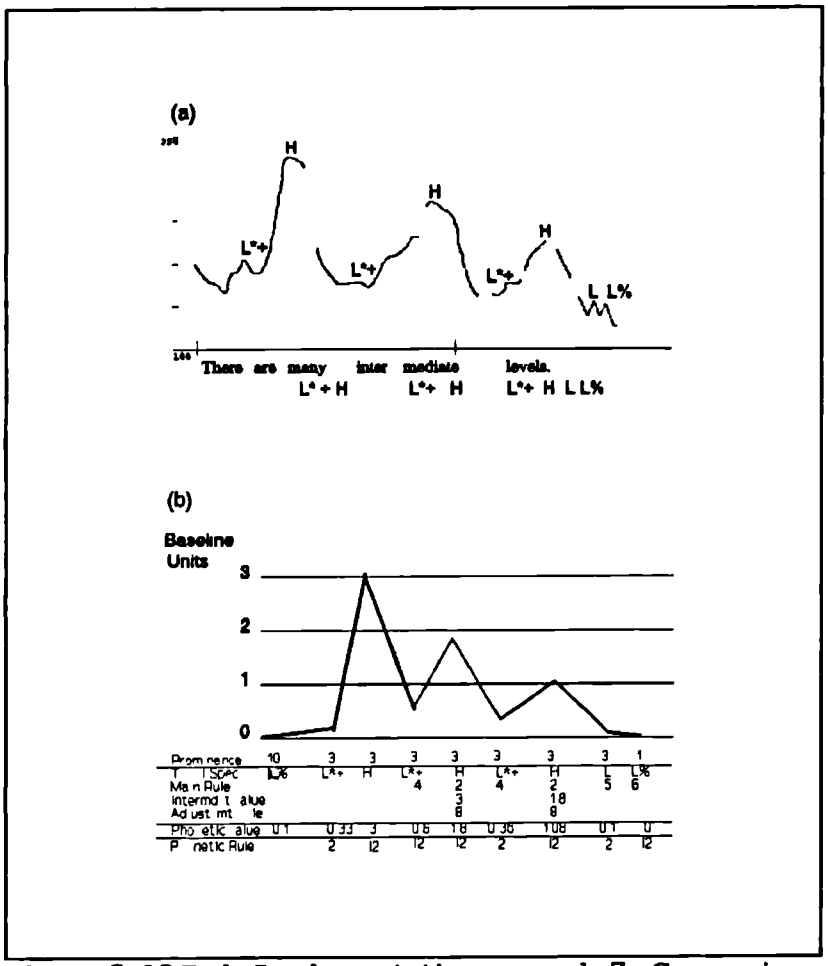


Figure 3.18 Rule Implementation example 7. Comparison of natural contour with contour generated by rule: (a) Adapted from Pierrehumbert 1980, Fig. 4.3, p.330; (b) Stylised contour generated by rules, with synopsis of rules applied in derivation.

Rule Implementation Example 8 (Fig. 3.19)

The last downstep example consists of a sequence of L+H* bitonal pitch accents. Tonal value assignment is according to rules in exactly the same way as in the previous example.

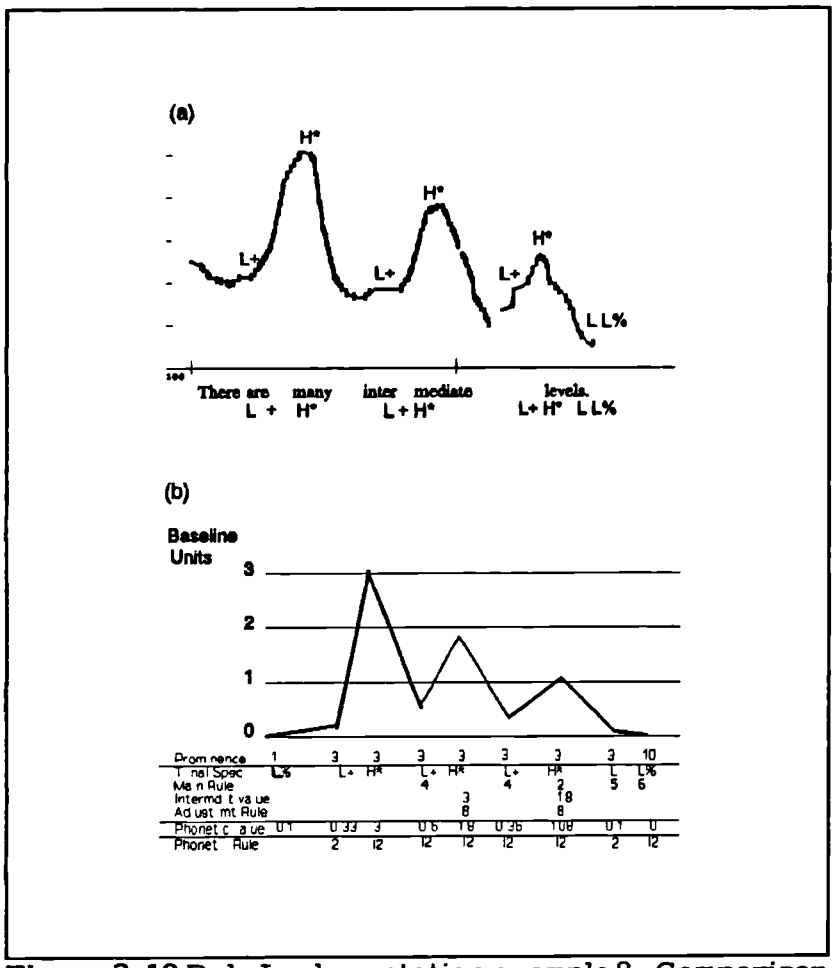


Figure 3.19 Rule Implementation example 8. Comparison of natural contour with contour generated by rule: (a) Adapted from Pierrehumbert 1980, Fig. 4.4, p.330; (b) Stylised contour generated by rules, with synopsis of rules applied in derivation.

The whole set of rules, and not just those directly involved in downstep, has been given, because it is the process of downstep whose example is used to motivate the form of rules; in particular, the fact that they are all left-to-right local-context-sensitive in application. This fact is explicitly stated by Pierrehumbert:

"We will assume that the series of phonetic values of the tonal sequence is initialized by the speaker's choice of the first H tone, for expressive purposes. Given that the value of a downstepped H is lowered by a factor of k relative to the preceding H, a chain of downstepped H's then results in a sequence of phonetic values of the form $k/H_1/$ $k^2/H_1/....k^n/H_1/$. [...] We

will see ... that this general approach to tonal evaluation can be motivated for the other rules which evaluate tones. That is, once the sequence of tonal values for the phrase is initialized, the value for each new tone, T_{i+1}, is computed as a function of its prominence and of the phonetic and phonological values of tones to the left." (Pierrehumbert, 1980, pp50-1).

Furthermore, computational details of nearly all rules are fixed by the requirement of the incorporation of downstep, even extending down to the existence of two forms of interpolation rule. This is a result of Pierrehumbert's attempt to constrain the power of the finite-state grammar generating sequences of tones in a first-order transition network; without the involvement of downstep, and given that the finite-state grammar doesn't allow the observation of non-local dependencies in generating contours, many wildly varying contours would in principle be generable. Before seeing what contours could be generated in these circumstances, it is necessary to see what constitutes the non-involvement of downstep, which will at the same time show how much the rules are involved in the process.

In the first place, not to take account of downstep would involve elimination of the downstep adjustment rule. The absence of that rule would imediately remove the need for the nonmonotonic 'sagging' interpolation rule between H tones, the requirement for which Pierrehumbert had noted as an irritation (1980, pp. 70-71). She had preferred to account for the dip between peaks in such contours (as seen in Fig. 3.15) with the floating L tone of a bitonal (H*+L or L+H*) pitch accent, but notes two problems.The first is that that would have meant that one of the two downstepping sequences involving such H+L accents (as seen in Figs. 3.16 and 3.17) would then have had to be accounted for by a sequence of monotonal H* accents. Thus, if we were to assume that the H*+L accent is used to account for contours like that in Fig.3.15, then it could not be used to account for contours like that in Fig. 3.17[13]. A sequence of H* accents would then have to be used to account for that contour, using the remaining interpolation rule (I2 in the synopsis above) to account for the total form of the contour. But then there would be a new context for downstep, and the form of the downstep rule would be considerably complicated. The second problem is that it would, in her words,

---

[13] This is because of the absence of sagging interpolation between the accents in Fig.3.17.

"make the rule an exceptional one cross-linguistically, since downstep is ordinarily found in sequence with alternating tonal types" (Pierrehumbert 1980, p.71).

Obviously, the elimination of the downstep rule nullifies those problems, and so the need for the nonmonotonic interpolation rule on those grounds is done away with[14]. However, there are other grounds for its retention, viz. that it accounts for the fact that when there is not enough segmental material between two peaks in a contour, no dip occurs. The form of the two-piece polynomial interpolation function between two H* peaks can be chosen to conform to this behaviour, whereas it would seem that a special rule of L tone target undershoot is required to account for it in interpolation between the tones in the sequence H*+L H(+T). In fact, that problem also disappears (at least for Pierrehumbert's 1980 thesis) , when it is considered that such sequences can always be considered to be of the form H*+H H(+T)[15]. For there would be no prima facie evidence that requires there to be a particular tonal sequence accounting for a particular sequence of peak F0 values in a system where underlying prominence values more directly determined those F0 values, which would be the case without a downstep rule.

In the second place, the set of rules computing values for L tones appears to have been set up primarily to accord with contour behaviour in downstep contexts. Rule 10 sets the value of the L tone in a H+L* pitch accents as the phonetic value of the H tone adjusted by the downstep factor k. This enables the overt stepping pattern seen in contours like that in Fig. 3.16 to be generated. Pierrehumbert claims that it also accounts for local behaviour in

---

[14] In fact, the second problem had already been obviated by Beckman and Pierrehumbert's (1986) claim that downstep is triggered by any bitonal pitch accent, without the requirement that there be an alternating tonal sequence.

[15] So, when the distance between the outer two H tones is large enough for a dip, and a dip occurs, the contour could be analysed as H*+L H. When no dip occurs over the same duration, the contour could be analysed as H*+H H; and when there is no space for a dip, it could be analysed as H*+H H or as H* H. Under this interpretation, there would be a phonological difference between the short contour comprising two peaks without a dip between them and the long contour comprising two peaks with a dip between them; Pierrehumbert, on the other hand, treats these as identical phonologically (1980, p.71). This difference of interpretation stems directly from Pierrehumbert's observation of the primacy of static tones as phonological entities over pitch movements.

other contexts, and a scan through the appendix of her thesis reveals three examples of similar nature which do include H+L* accents but where the iterative downstepping pattern doesn't occur. These can be seen in Figures 3.20, 3.21 and 3.22

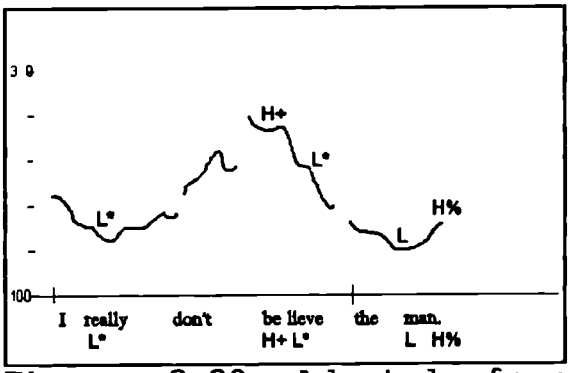


Figure 3.20 Adapted from Pierrehumbert 1980, Fig.4.20, p.342.

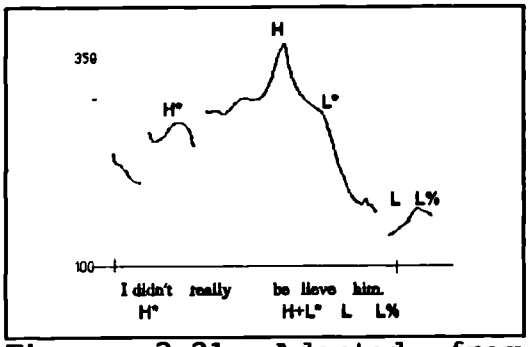


Figure 3.21 Adapted from Pierrehumbert 1980, Fig.4.32, p.350.

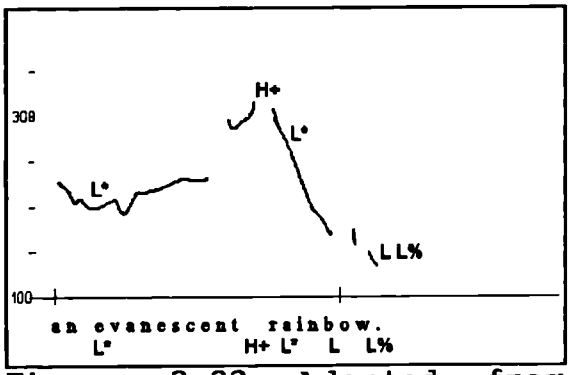


Figure 3.22 Adapted from Pierehumbert 1980, Fig.4.48, p.365.

In all three cases, a rising head is terminated by a falling or falling-rising nucleus, where the fall begins from a point below the peak value reached in the head. In the first case, the head starts low, and the contour is analysed as L* H+L* L H%. In the second case, the contour starts relatively high, and the contour is analysed as H* H+L* L L%; in the third, it starts low, and the contour is analysed as L* H+L* L L%. However, it is possible to discount these analyses, on the following grounds: under all three analyses, it should be possible to extend the duration (up to a reasonable limit) of the syllable associated with the L* tone of the bitonal pitch accent. This is because the stressed syllable should in principle be able to bear varying degrees of the concomitants of accent other than pitch obtrusion, that is, duration and amplitude; the principle is embodied in Anderson et al's (1984) implementation of Pierrehumbert's

system, in which associated tones are given an inherent, in principle variable[16], duration, but unassociated (floating) tones, such as the H+ tone in these examples, are not, and are just interpolated to and from. That being the case, it should be possible for the contours in the three examples to be produced by a native speaker with a prolonged level stretch at the point of the L* tone in the bitonal pitch accent. To the author, these are difficult to produce. There appears a natural tendency to produce a dynamic contour (a fall of whatever slope) whose starting point is at the level of the phonetic value of the L* tone, rather than a level tone, in these contexts. In addition, contours with level tones in those positions sound odd. Admittedly, this is more the case for the example in Fig. 3.20 and 3,22 than that in Fig.3.21. That can perhaps be explained from a perceptual point of view as resulting from the fact that the rising H* H+ sequence in Fig. 3.21 sounds quite similar to a high prehead or head in that position, in which case a level tone at the position of the L* in the bitonal accent would sound much more acceptable; in fact, it could be argued that the rising sequence is simply a variant of the high level sequence, resulting from delayed target achievement. Thus, the evidence for the H+L* tone appearing in other than downstep contexts is not conclusive, and it appears that in the absence of a downstep rule, Rule 10 could be done away with without any adverse side-effects.

To continue with an analysis of the dispensibility of the rules computing L tone values once the downstep rule is eliminated, it is possible to see that Rule 4 has been specifically designed for accounting for cases of what Pierrehumbert refers to as partial downstep, such as appear in Figs 3.18-19, in contrast to the cases of total downstep, such as appears in Fig. 3.16. Her conception of this distinction, which is again adopted from analysis of African tone languages, is criticised below. Here, it is possible to see that the distinction is maintained only once certain assumptions are made about prominence relations, and also, and more importantly for the point being made here, that there is no need for the rule in the absence of the downstep

[16] The variability of associated tones is made explicit in Silverman's (1987, Ch. 5) reworking of Anderson et al's implementation; the potential for doing so within the same framework (in this case for the Japanese language) is also acknowledged in Pierrehumbert and Beckman 1988 (pp. 175-6).

rule, and that, in fact, it predicts wrong things about the shapes of contours in non-downstep contexts, as discussed below.

To see the former point, it must first be pointed out that Pierrehumbert views partial downstep as comprising those cases where the L tone in a H L+H sequence is lower than the final +H tone, which is downstepped (this can also be the case in a H+L H sequence as the rules stand, if the prominence of the first H tone is less than the second; generally, though, the prominence values on tones in downstep sequences need to be the same (see below), and in those conditions, the H L+H sequence is the relevant case. Now this could be achieved using a rule of the form

In H (+T) L+ : /L/ = n /H/ where 0<n<k

which is the same as Rule 4, without the inclusion of the factor Prominence(H)/Prominence(L) . The reason Pierrehumbert complicates the rule with that prominence ratio appears to be that she feels that it is not enough for n to be less than k in the typical H L+H downstep contexts; it has to be quite a bit less than k, in fact a value that makes L pretty near the baseline. Yet she has independent evidence that how near 'pretty near' is depends on the relative prominence on the L tone, a fact that is reflected in the prominence ratio of Rule 7, which, in terms of the sequence of tones, is the inverse of the prominence ratio used in rule 2. Consequently, it would seem, she deems it appropriate to incorporate an equivalently inverse ratio as an adjustment factor in Rule 4.

The only trouble is, it doesn't work for the sequence of tones comprising an H followed by an L unless some restrictions are made on the prominence of the respective pitch accents. In particular, Prominence(H) must be less than or equal to Prominence(L) if the correct results are to emerge. In addition, the value of n should be considerably smaller than k. To see why this is so, consider the following cases:

Case 1

n=0.5, k=0.6, Prominence(H_i)=Prominence(L)=Prominence(H_{i+1})

Then, in the sequence H_i L+H_{i+1}, if H_i has the value 2, L has the value 0.5*2*1 = 1, and H_{i+1} has the value 1.2 . But these are not the kind of phonetic values that reflect the contour patterns seen in Figs. 3.41-2. So the value of n must be reduced.

Case 2

n=0.2, k=0.6, Prominence(H_i)=Prominence(L)=Prominence(H_{i+1})
Then, in H_i L+H_{i+1}, if H_i = 2, L=0.2*2*1 = 0.4, and H_{i+1} = 1.2. These values give a better approximation to the contour patterns aimed for.

Case 3

n=0.2, k=0.6, Prominence(H_i)=1.5, Prominence(L)=Prominence(H_{i+1})=2
Then, in H_i L+H_{i+1}, if H_i =1.5 (here we assume that the phonetic value of H_i has not been adjusted by some application of downstep on it or prior to it), L=0.2*1.5*1.5/2 = 0.3, and H_{i+1} = 0.8. This is again a good approximation to the contour patterns aimed for.

Case 4

n=0.2, k=0.6, Prominence(H_i)=2, Prominence(L)=Prominence(H_{i+1})=1.5
Then, in H_i L+H_{i+1}, if H_i=2 (with the same assumption as in Case 3), L=0.2*2*2/1.5 = 0.53, and H_{i+1} = 0.45. Here, the contour is contrary to that aimed for, with the L higher than the second H.

Case 5

n=0.2, k=0.6, Prominence(H_i)=3, Prominence(L)=Prominence(H_{i+1})=1.5
Then, in H_i L+H_{i+1}, if H_i=3 (same assumptions as before), L=0.2*3*3/1.5=1.2, and H_{i+1} = 0.3. Again, this is quite contrary to the contour patterns aimed for, and gives an incorrect assignment of values to the L+H pitch accent.

In addition, there is a corollary of this situation, viz. that given that the value of H_{i+1} is less than the value of L in cases 4 and 5, but has been greater than it in previous cases, it is possible to imagine a set of parameter and prominence values in which L=H_{i+1}, that is, what would be considered a case of 'total downstep' (see below).

The correct relationship in cases 4 and 5 could be considered to be salvageable by making n even smaller, but it is clear that the situation is unstable; there is always the possibility of a sequence of prominence values of sufficently large differential value arising to create the circumstances in which L>=H in a L+H pitch accent. Thus there must always be a constraint on

the prominence values in a downstep sequence containing L+H pitch accents that no pitch accent has greater prominence than its successor[17].

It should be noted that the problem just detailed would still arise in the absence of the downstep rule. In that situation, in Case 5, L=1.2 and $H_{i,i}$=0.5. Thus Rule 4 is superfluous, and even harmful, to a rule set without the rule of downstep, and has been shown to be of the form it is because of particular expectations about downstep sequences. It could thus be dispensed with without any loss in predictive power. Indeed, the inverse scaling relationship between prominence and phonetic value would be better maintained by an adjusted version of Rule 7, as is seen below.

The next rule to be considered, Rule 5, determines the value of a L phrase accent as a fraction p of a preceding H tone. It thus has specific application and is not worthy of extensive examination here. It can merely be noted that in the absence of a downstep rule, such a rule would still be necessary, but there would be no reason to constrain the value p to be less than a downstep factor k. The constraint in fact indicates that the rule has been formulated to accord with L phrase accent behaviour in downstep sequences. There are certain circumstances in which it might be appropriate to have higher L phrase accent values than are predicted by the rules. House (1989) gave evidence of such values in utterances with falling nuclei but long tails.

Rule 6 would similarly still be required in a rule set without downstep. It is the only one of the main phonological rules whose form is not constrained by the existence of the downstep rule.

The form of Rule 7 is also constrained by the existence of the downstep rule, though more indirectly than the other rules computing values for L tones. Before discussing it, it should first be reiterated that we must assume that

---

[17] This constraint could perhaps be quite acceptable to Pierrehumbert; it accords with one of her expectations of downstep sequences, that the last pitch accent in such a sequence has greatest prominence. But observations below indicate that this expectation is unwarranted; and in any case, there are likely to be many cases in which the first accented syllable of a tone-unit (the 'head' accent) has less prominence than the nucleus, but more than its successor. That set of relations ($|A_{m+1}| < |A_m| < |A_n|$, where A_i = the ith Accented syllable, '|' indicates prominence, and $n > m+1$) could also be true of any sequence within the the tone-unit.

there is a direct relationship between the prominence of an L* accent which is the initial accent in a phrase and its phonetic value, such that the phonetic value is the reciprocal of the prominence. This is the only relationship which makes sense of the claims by Pierrehumbert that "the value of the first pitch accent in the phrase is a free choice, governed by pragmatic or expressive factors (1980, p. 144)" and that "the phonetic value of L* decreases if its prominence is increased" (1980, p.68) while "[t]he lowering of L's under prominence is .. prone to saturation" (1980, p.69). Now, Rule 7 predicts that the phonetic value of an L tone (be it a phrase accent or part of a pitch accent) following a L* pitch accent is computed as the product of the phonetic value of that L* pitch accent tone and the reciprocal of the ratio of its prominence to that of the L* pitch accent tone. This might seem to be a relationship between adjacent tones which is quite independent of downstep processes, and in a sense it is. That is, if the relationship between the prominence of the preceding L* pitch accent and its phonetic value were arbitrary, then the rule would scale the phonetic value of the L tone in a non-trivial way as a function of the two prominence values. Similarly, if that relationship were not direct, so that the phonetic value of the preceding L* tone was not the reciprocal of its prominence, as suggested above is required for an initial L* accent in a phrase, but was the result of a prior adjustment rule, such that the phonetic value was less than the reciprocal of its prominence, then again there would be a non-trivial computation of phonetic value involving the two prominence values. However, the only adjustment rules which are relevant to this context in the system of rules are Rules 10 and 4, the necessity for whose existence has been seen to depend on the existence of the downstep rule. In that rule's absence, the picture changes. That is, with no possible prior adjustment rule, then it would always be the case that the phonetic value of a preceding L* tone would equal the reciprocal of its prominence. In that case, Rule 7 could be interpreted as follows:

$$\text{In } L^*_i\ L_{i+1}: \quad /L_{i+1}/ = \frac{1}{\text{Prominence}(L^*_i)} \cdot \frac{\text{Prominence}(L^*_i)}{\text{Prominence}(L_{i+1})}$$

$$= \frac{1}{\text{Prominence}(L_{i+1})}$$

Thus, Rule 7 would only state the convention already taken to hold, that the phonetic value of a L pitch accent (or its associated phrase accent) is equal to the reciprocal of its prominence. Thus, the form and raison d'etre of Rule 7 is seen to be predicated on the existence of the downstep rule, whose elimination would also lead to Rule 7 being scrapped. At this point, it should be noted that in the view of the author, if a sound case could be made for consistently declining behaviour of L tones in H L+H downstepping sequences (and allowing the existence of the downstep rule), then an amended version of Rule 7 allowing for intervening +H and H+ tones would be more satisfactory within Pierrehumbert's general scheme than Rule 4 has been found to be. Perhaps Pierrehumbert realised that such a case could not be made, a problem that will be addressed below.

In the third place, the rule determining the phonetic values of H tones can be seen to be justified by the existence of the downstep rule. Here, an argument exactly parallel to that put forward to show the dependence of Rule 7 on the downstep rule can be made. Thus, on the assumption that the phonetic value of an initial H* tone equals its prominence, and that there is no downstep rule (nor any other adjustment rule applicable in the context of Rule 2), then the ratio of phonetic value to prominence of any H tone in a sequence is equal to 1, and so Rule 2 reduces to the already stated convention that the phonetic value of an H tone equals its prominence.

Thus, of all the rules presented, only Rule 5 (in an adjusted form), Rule 6, Rule 9 (the upstep rule[18]), Rule I1 and the spreading rules are seen not to depend on the existence of the downstep rule. Of the main phonological and phonetic rules, those that would remain upon the elimination of the downstep rule relate only to the phrase accent and boundary tone. All other tones would have their phonetic values computed direct from their prominence values. This happens to be the case already for the tones in pitch accents in certain contexts; the value of the first pitch accent in a phrase has been seen to be a free choice. We can only assume (and have assumed in the examples in Figs. 3.12-3.19) that the same is true of an initial boundary tone. In fact,

---

[18] Even this rule relies for satisfactory results in some of its applicable cases to the existence of the downstep rule. That is, when the H phrase accent is preceded by a H*+L pitch accent, it is downstepped, which allows for the correct generation of H*+L H L% and H*+L H H% vocative contours.

there are some contours in which no context for downstep exists, and as a result, in the absence of other adjustment rules, the phonetic values of the tones comes directly from the underlying prominence values, except for the phrase accent and final boundary tone. This is the case in Fig. 3.23, which shows an example of what has been dubbed the surprise-redundancy contour[19] (because it can be used equally in contexts in

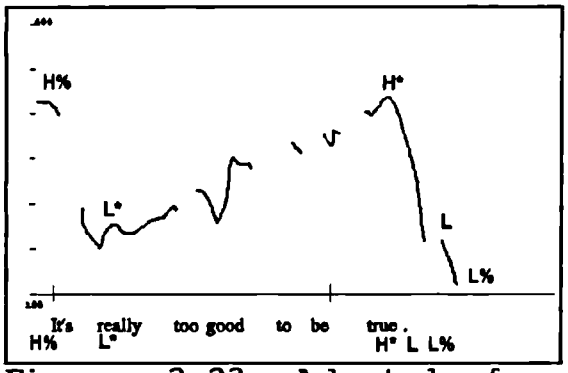


Figure 3.23 Adapted from Pierrehumbert 1980, Fig. 4.29, p.348.

which the speaker expresses surprise and contexts when the speaker lets it be known, perhaps with indignation, that the expressed piece of information is really redundant and should have been known by the hearer). The initial boundary tone is given a free choice, the initial L* pitch accent too, and because no context fits the following H* tone in the Rule Set, it too has to be given a free choice of phonetic value (directly from the free choice of prominence). Only the L phrase accent and L% boundary tone are given values computed by rule. In these circumstances, there is no need to implement the left-to-right local application of tonal sequence rules for the majority of tones. The values could be assigned in any order, without affecting the results. Yet, since that is the case, there is no clear argument for suggesting that the rules should be local ones applied to a first order transition network. They could equally apply in global fashion to all tones in the sequence at once.

Perhaps Pierrehumbert could accept the existence of global rules in the computation of certain iconic forms such as the surprise-redundancy contour. Indeed, it might be considered a strength of her model that the possibility of such global computation arises in exactly the case in which an iconic contour is modelled by a sequence of tones none of which fits the context of the local rules. However, it is clear she would not accept globality across the board, yet we have seen that in the absence of the downstep rule,

[19]It was assigned this meaning by Liberman and Sag (1974)

the possibility for global application would be high, because most of the local rules would have disappeared. Thus it can be seen that not only does the incorporation of downstep into an account of English intonation (a) allow for the specification of only two pitch levels in the intonation system, (b) secure the distinction between monotonal and bitonal pitch accents, and (c) account partially for the use of the phrase accent through its use in analysis of the form of 'stylised' intonation forms such as the calling contour, but (d) it also crucially underpins the substance of her set of intonation rules, that is, their left to right application and the absence of non-local dependencies between tones.

The use of some form of downstep rule as an essential part of the intonational description of a language has become commonplace since its use in Pierrehumbert's thesis. Before assessing how essential the downstep rule is in such a description, it is appropriate at this point to consider another approach to the description of English intonation which incorporates an account of downstep but which is not constrained to having only local dependencies between tones. This will demonstrate that a downstep rule does not force the use of a finite state grammar for generating intonation contours, but only favours such a choice, because of the locally iterative nature of its application.
The approach is that of Ladd (1990).

3.3 LADD'S ACCOUNT OF DOWNSTEP

Ladd (1990) presents an analysis of intonational structure in English which attempts to constrain even further than Pierrehumbert the form of intonation contours in respect of the relative heights of adjacent (and non-adjacent) tones, by requiring them to conform to some metrical structure. At the same time, variations in prominence are restricted in occurrence to the domain of a register, whose upper and lower bounds are maintained for a variable amount of time, depending on the configuration of the associated metrical structure.

Ladd's inventory of tonal forms is more restricted than Pierrehumbert's, comprising just the forms H, L, HL, LH and the boundary tones L% and H%. In Ladd (1983), he justifies this economy in the set of primitives by pointing out that it allows more generalisations to be made about particular

adjustments that are made to basic tonal forms which are independent of the tonal form; at the same time, these adjustments, which are mediated by tonal features, account for the variety of contour shapes which Pierrehumbert generates through her richer tonal inventory. For instance, Pierrehumbert's L*+H pitch accent is just Ladd's H accent with the application of the feature [+delayed peak] [20], but the same feature could be applied in principle to a nuclear LH accent (he allows LH accents only in nuclear position) to yield the delayed final rising sections typical of English nuclear rises.

In Ladd's latest phonological model (the phonetic details are actually implemented on computer at CSTR in Edinburgh), the tonal forms constitute terminal elements of a binary branching (metrical) tree, where non-terminal binary branches at the same level terminate with an h and an l node. This will be examined in detail in section 3.3.1 below.

3.3.1 Metrical structure and Downstep

The functional load of the [downstep] feature in his 1983 account has been taken up by a particular rule set which applies to nodes in the metrical tree to determine whether the upper and lower bounds of an operative register remain the same or are 'downstepped'. The output of this rule set is a string of tones with interspersed commands for shifting the register down (in fact, incrementing an index i, as appears in the rules below). First, the Highest Terminal Element (HTE) of the tree has to be defined, as follows:

'The HTE of any metrical tree or subtree is that terminal element arrived at by following all the branches labeled h from the root of the tree or subtree'.(Ladd 1992, p.44)

---

[20] Pierrehumbert and Beckman (1986,pp259-60) in fact object to this characterisation of the forms they represent with a L*+H pitch accent, claiming that it doesn't account sufficiently for the particular phonetic detail of the rise-fall pitch accents in question. Indeed, their objection seems to be justified, as the following diagrams (3.25) adapted from Ladd's paper indicate that his H accents rise from and fall to the medial reference line in a register, and not to the baseline, whereas Pierrehumbert's L*+H accent involves a truly low dip before the peak.

It must be assumed from the example given (discussed below) that this rule applies to terminal nodes of the tree as a null case, so that the tonal elements attached to the terminal nodes can be HTE's even if no branches are labelled h (because the only branch is labelled l). That is, in the tree in 3.24 below,

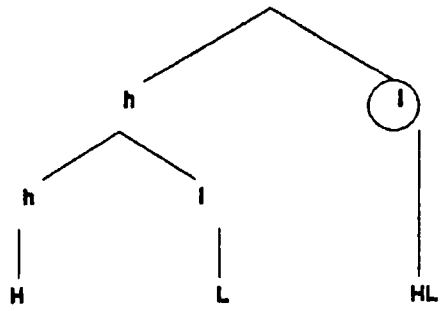


Figure 3.24 An example of Ladd's metrical tree structure used for the scaling of tones.

the subtree dominated by the (circled) l node to the right of the tree has an HTE of the tonal form HL, even though no h nodes are passed through.

Next, Ladd formulates a (first approximation to a) Relative Height Projection Rule (RHPR) :

'In any metrical tree or constituent, the HTE of the subconstituent dominated by l is:

(a) One register step lower than the HTE of the subconstituent dominated by h when the l subonstituent is on the right;
(b) at the same register as the HTE of the subconstituent dominated by h, when the l subconstituent is on the left.' (Ladd 1990, p.44).

This needs to be augmented by a convention that the HTE of a subconstituent always determines the upper bound of a register associated with that subconstituent. Thus, case (a) constitutes a case of register downstep.

The phonetic values of the target values for the tones are determined by the following rules:

' $\log(F0) = \log F_{min} + f(N) \times f(T)$

where F_{min} is the speaker baseline, f(N) is the current register setting, and f(T) is the current tonal specification' (Ladd, 1992, pp53-4).

' $f(N) = N \times d^{i}$

where N is the speaker specific range parameter that specifies the default initial register...[, d is the] factor by which register steps down...[and] i is a positive integer that increments by 1 at each downstep' (Ladd 1992, p54).

' $f(T) = w^{T \cdot p}$

where T = +1 for H, -1 for L, and 0 for the middle of the register. The value of p (for prominence) must lie between 0 and 1. If the difference between reduced and nonreduced prenuclear prominence turns out to be a categorical rather than a gradient difference [..] then p would have only two values, 1 and some fraction.' (Ladd 1990, p.54).

(Note that in the above, d (which amounts to a downstep constant similar in effect to Pierrehumbert's k) is given a value of 0.8, and w (an indicator of register width) a value of 1.5 at CSTR.)

The significance of that last comment about permissible values for p is that the nuclear tone (where the domain of the nuclear tone appears to be the register[21], a purely intonational entity which would appear to map onto the prosodic entity the Tone Group (TG), which is recursively definable within the domain of a major phrase in Ladd 1986) always bears maximal prominence within a particular register stretch. This means that, in the absence of any exception rules allowing transgression of the upper bound of a register because of local prominence, the final tone in a register stretch will always

---

[21] Ladd does not explicitly state this to be the case; in fact, some of his registers contain only H tones, and in Ladd 1983 these are only allowed prenuclearly, nuclei being reserved for HL or LH tones. However, there is an underlying thrust in his work which involves extending the concept of nuclearity to lower-level domains in order to account for the scaling of tones, in direct contrast with Pierrehumbert, who largely dispenses with the concept of nuclearity.

be the highest. Two examples of the output of the rules are given in Fig. 3.25.

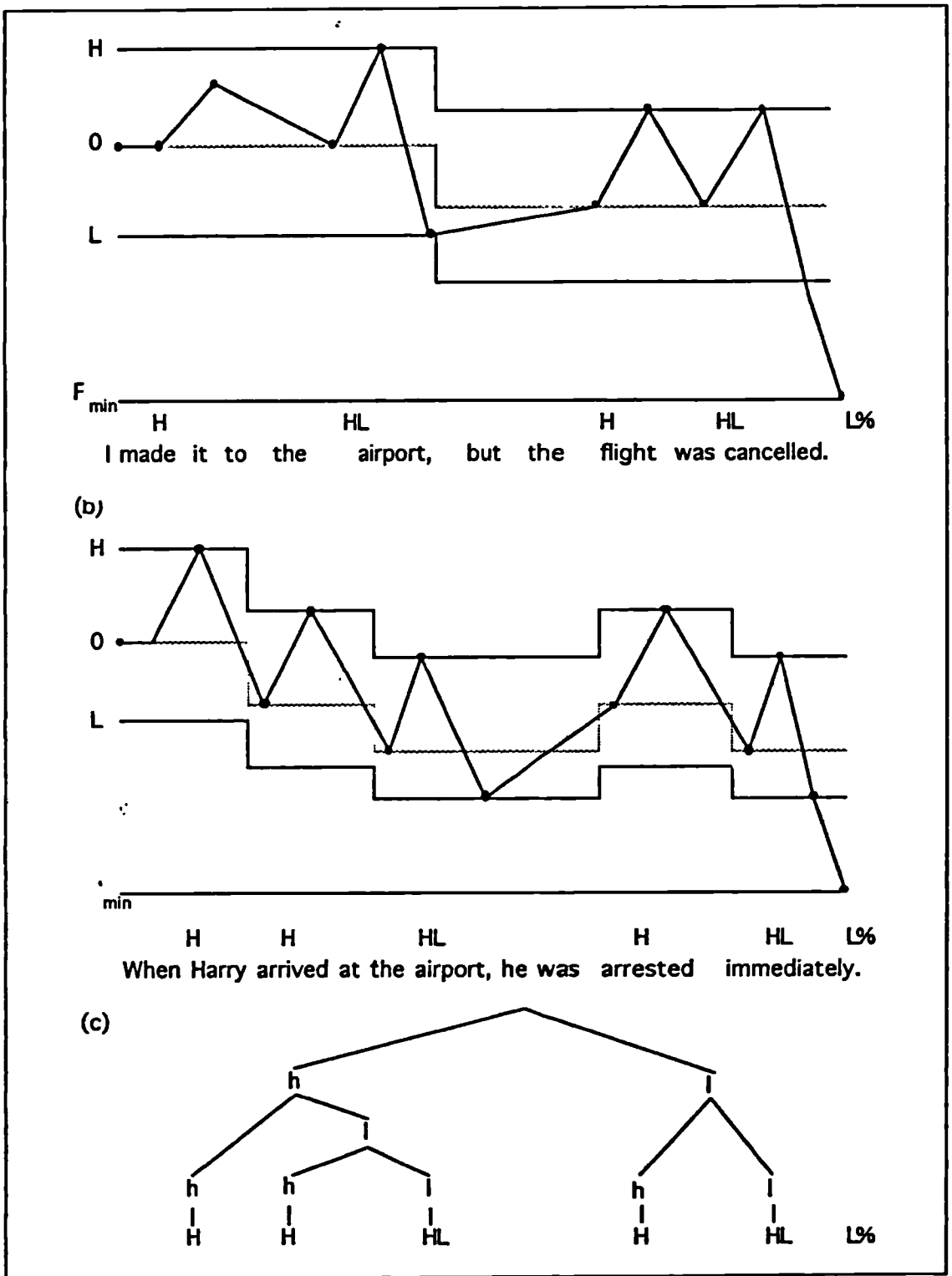


Figure 3.25 Two examples of output of Ladd's rules, taken from Ladd, 1992, p40. The tree in (c) applies to the contour in (b) only.

The contour produced in (a) is generated in the following sort of way(the value of F_{min} is taken to be constant (70Hz) here, though Ladd suggests it could gradually decline through the course of an utterance; this would constitute 'true declination', as opposed to downstep) :

Assuming a value of N = 1, F_{min} = 70Hz, and T= 0 when T is neither H nor L, and the other parameters in the rules as indicated above, the initial boundary tone (assumed in the figure) could be given a value of

$$\exp(\log(70)+1*0.8^{0} * 1.5^{0*0.5}) = \exp(\log(70)+1) = 190\text{Hz}.$$

(whether it is or not depends on how Ladd treats boundary tones).
The following H tone is given a value of

$$\exp(\log(70)+1*0.8^{0} * 1.5^{1*0.75}) = \exp(\log(70)+1.5) = 271\text{Hz}.$$

The H of the following HL pitch accent is given a value of

$$\exp(\log(70)+1*0.8^{0} * 1.5^{1*1}) = \exp(\log(70)+1.5) = 314\text{Hz}.$$

The L of that same pitch accent is given a value of

$$\exp(\log(70)+1*0.8^{0} * 1.5^{-1*1}) = \exp(\log(70)+0.667) = 136\text{Hz}$$

Next, there occurs a register shift through downstep (i is incremented to 1). Then, the H tone of the first H pitch accent in the shifted register is given a value of

$$\exp(\log(70)+1*0.8^{1} * 1.5^{1*1}) = \exp(\log(70)+1.2) = 232\text{Hz}$$

The H tone of the following HL pitch accent is given the same value (232Hz), because it has the same prominence as its predecessor. The L tone of that pitch accent is given a value of

$$\exp(\log(70)+1*0.8^{1} * 1.5^{-1*1}) = \exp(\log(70)+0.533) = 119\text{Hz}$$

Finally, the L% boundary tone is given a value of F_{min}.

This example underlines the point just made, that the 'nuclear' tone - the rightmost pitch accent - in a domain coterminous with a register stretch, has maximal prominence, i.e. p=1.

The next example, in (b), has more register shifts during an utterance of equivalent duration. Each shift of the register corresponds to the initiation of a new (minor) tone group, so each shifted register contains a nuclear tone. The register shifts are caused by h-l sequences in subconstituents of the metrical tree associated with the text, which appears in (c). In the first clause of the sentence, there are two downsteps, corresponding to the h-l

sequence appearing at two levels within the subtree dominated by the topmost h. At the beginning of the second clause, there is another downstep, but the shift down is relative to the register whose upper bound is determined by the HTE of the subtree dominated by the topmost h (according to the RHPR clause (a)), that is the first register set in the utterance. So, relative to the preceding register (which is two shifts down) there is a shift up. Finally, there is a further register shift down, corresponding to the h-l sequence in the subtree dominated by the topmost l.

3.3.2 Discussion

Despite the hierarchical structure which determines downstep domains, there are some similarities with Pierrehumbert's account of the process of downstep. In fact, within a given downstep domain, (that is, within a stretch containing no downstep rule application), the computation of F0 values on accented syllables depends largely on the variable of prominence, with the binary variable of tonal type obviously determining the location within a register that a tone is centred. However, Ladd imposes more constraints on the relative prominence values on consecutive accented syllables.

Futhermore, there are global dependencies involving downstep in Ladd's model, and these are of two types: (i) within a subtree, the use of the index i means that global information is passed to the function f(N), viz. the number of downsteps that have occurred within that domain. As a matter of fact, this information is not used other than directly in f(N), and the index mechanism could be replaced by local multiplication by the downstep constant. (ii) More importantly, there is a look-back mechanism over an arbitrary number of intervening tones to determine the height of a register which has just been shifted. The number of tones has to be arbitrary because a downward shift in register can occur as the result of an h-l sequence at any level in the metrical tree. Thus, the imposition of the hierarchical superstructure of the metrical tree provides much information about the tonal structure of the utterances as a whole, so that in the process of computing the F0 value of the next tone in the terminal string of tones, any register shift doesn't just use local context, but higher-level context, in order that the correct value of shift be applied[22].

---

[22] Of course, within this hierarchical recursive structure, each individual register shift, as specified by the RHPR, is a purely local procedure.

What is of additional interest in Ladd's model is that he is also suggesting the existence of a look-forward mechanism, specifically within the domain of a register. This is indicated by his apparent requirement (and attempt to corroborate with various experimental evidence in his 1990 paper) that the rightmost tone in that domain (i.e. the nucleus) have maximal prominence within it. This means that preceding tones have to have their F0 values scaled relative to a following tone (which is also an arbitrary number of intervening tones forward). In addition, this look-forward mechanism only holds in the absence of downstep; in any sequence of pitch accents terminated by a unique nuclear tone, the subtree dominating it must contain l-h sequences at every level, for an h-l sequence would trigger downstep, which would shift the register and introduce a further 'nuclear' tone just prior to the point of register shift. Thus, there are two global referral mechanisms within the model, one looking forward to the prominence of a nuclear tone within a register, and secured by the existence of a l-h sequence in the dominating subtree, and one looking back across registers to the pitch height of a preceding register, and secured by a h-l sequence in the dominating subtree.*

Ladd's recourse to the concept of the nucleus (which he has consistently had in his intonational research since Ladd 1980) is in direct conflict with Pierrehumbert's approach in Pierrehumbert 1980 which denies the rightmost pitch accent in an intonational phrase special status. This is compatible with the following purpose: it negates the possibility of look-ahead taking place; but more importantly, it secures one of the main motivations for her conception of downstep, that its existence accounts for the odd fact that the most prominent of the accents in an intonation phrase, the rightmost, has often very reduced pitch in a downstepping sequence; it is the linear local left-to-right application of downstep that achieves this result. We shall see in the following section that there are problems with this claim.

3.4 PROBLEMS WITH LOCAL LEFT-TO-RIGHT IMPLEMENTATION

In order to examine the problems with linear local left-to-right implementation, it is necessary to take a closer look at Pierrehumbert's 1980 model. Having seen that it is the process of downstep which enables her to constrain the intonation contours generated by her model in respect of the relative pitch heights of successive accents, whilst maintaining the discipline

* See footnote in addendum on p. 338

of left-to-right local implementation of rules, it is appropriate here to consider the types of sequence that would result if the downstep rule were done away with, so that the heights of successive pitch accents were unconstrained by intonation-specific phonological rules. It is necessary to be sure that some such constraints are needed; this will lead naturally on to a discussion of whether it is appropriate to use an adjustment rule specifically of downstep.

As a first example, consider the tonal sequence appearing in Fig.3.16

L%	H*	H+L*	H+L*	L	L%
3	3	3 3	3 3	3	3

with underlying prominence values

which could be given the following assignment of phonetic values:

0.33	3	3 0.33	3 0.33	0.33	0

This would yield a contour of the form:

There are many inter diate me le vels.

which is only marginally acceptable, according to the current author's judgment. This suggests that it would be the responsibility of that part of the grammar which assigns appropriate underlying values of prominence to the tonal elements to ensure acceptable intonational forms. For instance, a change in the underlying prominence of the first pitch accent in the above sequence could result in the following rather more acceptable sequence being generated:

	H%	H*	H+L*	H+	L*	L	L%
Prom:	1.5	1	3 3	3	3	3	1
P.val:	1.5	1	3 0.33	3	0.33	0.03	0

There are many inter diate me lev els

Yet at the same time, there would be nothing to prevent the following sequence of phonetic values being generated:

	H%	H*	H+L*	H+L*	L	L%
Prom:	0.5	2	1.5 1.5	2.5 2.5	2.5	1
P.val:	0.5	2	1.5 0.66	2.5 0.4	0.04	0

diate

ma ny in ter

a re

There me le vels

which could only be the output of a speaker who was extremely unsure of what message they were trying to convey, managing thus to fall between two stools.

How, then, would the sequence of prominence values be constrained? First of all, it has to be noted that Pierrehumbert's approach reflects a particular view of the prosodic structure of English as opposed to some other languages, that is, that English is primarily a stress-accent language whereas Japanese and Swedish, say, are pitch-accent languages; and that by this is meant that the primary means of accenting syllables (and thereby, words) in English and other stress-accent languages is by stressing them by means of relatively increased amplitude or duration, whereas in pitch-accent languages, it is by pitch marking. Corroboration for this view is that that state of affairs is reflected in the lexicon, in which, for non-reduced words (i.e. most content words such as nouns, verbs, adjectives, and so on, and many function words whose syllable peaks do not consist solely of reduced vowels), there is always a syllable marked which bears main word stress, independent of any considerations of variation in pitch. That contrasts with Japanese, according to Pierrehumbert and Beckman 1988, which marks accentuation in the lexicon by means of H(igh) and L(ow) tone sequences.

This view seems to fly in the face of a prevailing view of accentuation, at least for British English, that was generated in the light of experiments performed by Fry in 1958 (Fry 1958). In those experiments, it was found that

the most significant marker of stress (or perceived prominence) on a single syllable was pitch obtrusion, followed by duration and then amplitude. Beckman and Pierrehumbert (1986:271-2) argue, citing a number of references by way of experimental support, effectively that pitch obtrusion can be considered a sufficient condition for the perception of stress, and is in fact the most reliable correlate of perceived prominence, but that relatively high amplitude, long duration and associated lack of reduction of vowels are necessary conditions for it. This claim is borne out by the fact that in unaccentable syllables (as in post-nuclear position), that is, syllables which cannot have their prominence marked by significant pitch obtrusion, the markers of stress are exclusively amplitude and duration (Huss 1978)[23]. In accented syllables, they are augmented (but not displaced) by pitch obtrusion.

There is in fact no mismatch between the countervailing claims; the one claim (stemming on one side of the atlantic from Fry's experimental work, e.g. Fourcin 1962, but paralleled by auditory research in America by Bolinger, e.g. Bolinger 1965) is that relative pitch height of, or dynamicity of pitch movement on, accented syllables is the primary marker of perceived prominence; this is supported by the evidence of perceptual experiments. The other claim (from within the generative school of prosodic research, e.g. Beckman 1986, Beckman and Pierrehumbert 1986) is that relative duration and amplitude are the primary markers of perceived prominence; the consistency of the correlation between those factors and stress can be seen most markedly in production experiments, and it is only because of the masking effect of pitch obtrusion that such consistency is not also achieved in perception experiments; indeed, their primary effects can be discerned more clearly in locations where there is no pitch obtrusion.

At its crudest, the lack of mismatch between these claims can be shown by pointing out that one claim is based in a field of research whose primary orientation is perceptual or auditory, and the other in a field whose primary orientation is productive or synthetic. The one field explains productive aspects of speech communication in terms of perceptual targets, the other explains perceptual aspects of speech communication in terms of productive

---

[23] There is evidence that postnuclear stressed syllables can have significant F0 excursions, however; cf. Pierrehumbert herself (1980 ch.5).

targets (the 'motor theory' of speech perception, v. Liberman et. al. 1967). Thus, the perceptually most relevant concomitants of stress (i.e. pitch obtrusion) are considered primary by the former discussants, and the productively most relevant concomitants (i.e. amplitude and duration increases) are considered primary by the latter. Restricting the observations of each party to their proper domains would mean that the two bodies of observations could be incorporated into a more unified theory of speech communication, in which the various factors of stress production and perception could be given their appropriate priorities.

This may be considered a facile analysis of that debate, and indeed, more needs to be said. Those within the perceptual school can claim that, at base, you can't have production without perception, but you can have perception without production, which necessarily makes those aspects of communication which are more perceptually relevant primary in a unified theory; (the current author investigates the validity of this side of the argument to the extent that one of the experiments in this thesis studies the role of perception in controlling one aspect of speech communication, viz. declination in intonation.). Those within the generative tradition and propounders of the motor theory of speech perception (which streams of thought could be analysed as symbiotic)[24] could perhaps take a more abstract view, and claim that their positing of the primacy of productive processes is a claim about the organisation of the human speech communicative faculty in the brain.

---

[24]It is not suggested that the generative model of human communication is committed to the primacy of speech production by virtue of generating terminal representations of utterances; the author is aware that the term 'generative' refers to the study of the capacity of grammars of a language to generate all and only the sentences of that language. As has been noted above, generative analyses could be equally considered analyses of the language itself (i.e. competence models) and as such are, strictly speaking, equivocal about the question of the primacy of speech production and perception. It's just that when one gets down to the phonetic level, the computational machinery inherent in the generative paradigms combined with the need to take into account the productive and perceptual facts as elicited by the conduct of experiments tends to favour a view of the generative process as one mirroring actual productive processes. This state of affairs naturally leads to the idea that productive processes are primary in speech communication.

Returning to the particular point at issue, the current author thinks that in fact, neither theory of speech communication secures a result about the primacy or otherwise of pitch obtrusion in marking prominence. Sometimes, if the message imparted by pitch variation is of considerable importance, a speaker (of a stress-timed language such as English) will adjust the duration of a vowel so that the full form of a nuclear pitch configuration, say, is imparted (it is to be noted here that the duration is then dependent on the pitch variation, the point made in Bolinger 1965). On other occasions, if the lexical message is more important, the same speaker will rely on rhythm (as mediated by relative duration and, perhaps, amplitude) to mark prominence, and intonational forms can often be curtailed or even elided in such circumstances, and will be 'made to fit' around the rhythmic structure. On still other occasions, when there is little new information to impart, more fixed forms of intonation and duration will come into play, and the question of prominence will lose its relevance (as in stylised intonation contours - see Ladd 1978, and 1980, Johnson and Grice 1990, House and Youd 1991 - which mark peculiarly ritualistic episodes of speech communication). Indeed, it is speculated here that such occasions need not correspond to speech acts separated in time, but may co-occur within a single utterance, so that all the different markers of prominence gain primacy in quick succession (on separate accented syllables).

This view of things does, it must be said, relate to a model of intonation in speech communication which accounts for a speaker's or listener's performance. Yet even in models of intonational competence, there is no clear view on the primacy of pitch variation as a marker of prominence. Some (e.g. Selkirk 1984) have the tonal primitives as primary (and accordingly prior) in the generation of prosodic patterns. Others (e.g. Liberman 1975, also see Ladd 1992) have tonal primitives and the lexical concomitants of stress ('tune and text', in his, and Pierrehumbert's words) as parallel and of equal priority in the generation of prosodic patterns and conveyance of prominence. In Liberman 1975, the final patterns of prominence are determined by fitting tune and text trees to a metrical grid (developed further in Liberman and Prince 1977) which allows different assignments of prominence to be made on individual syllables as long as they are consistent with the relative prominence marked in the text tree. In 3.26 is an example

of such a tree (exactly parallel in form to the tonal trees discussed above in relation to Ladd's theory in Ladd 1992):

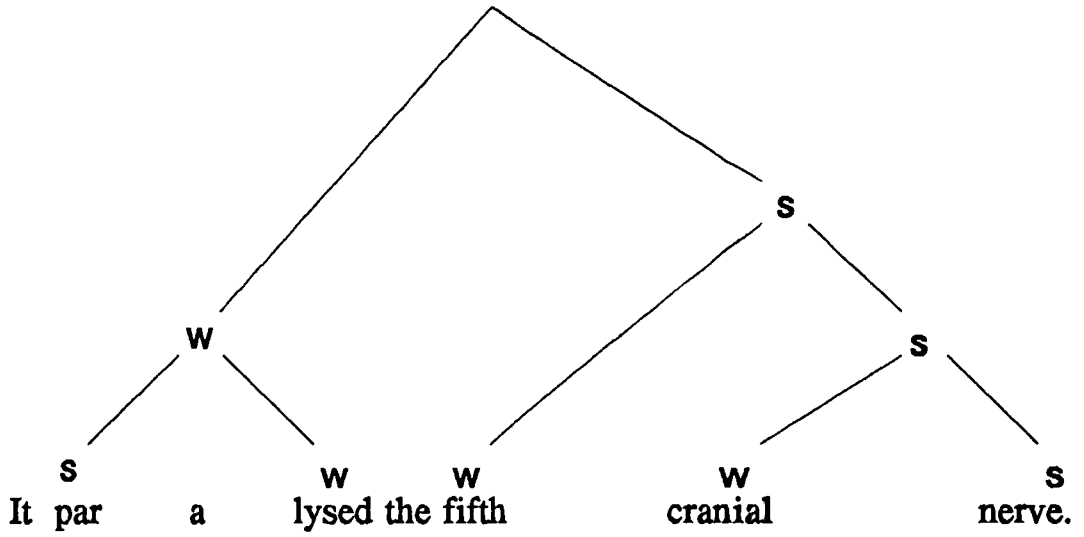


Figure 3.26 An example of the sort of text tree found in Liberman 1975

The nodes marked 's' and 'w' are strong and weak nodes respectively. Each subtree of this tree has a 'designated terminal element', which is the terminal element reached by passing through all the strong nodes of the subtree. The designated terminal element of the whole tree is the syllable 'nerve'; this syllable thus has the greatest metrical strength. The relative metrical strengths of all the terminal nodes in the tree are determined by the fitting of a metrical grid to the tree, which, after Liberman and Prince 1977, has to conform to the Relative Prominence Projection Rule (cf. The Relative Height Projection Rule in Ladd 1992)

Relative Prominence Projection Rule

In any constituent on which the strong-weak relation is defined, the designated terminal element of its strong subconstituent is metrically stronger than the designated terminal element of its weak subconstituent.

(Liberman and Prince, 1977, p.316)

Examples of metrical grids conforming to this rule in respect of the above tree are as follows (the grid examples are suggested in Pierrehumbert 1980:36-7, and include only main word stress) :

It paralysed the fifth cranial nerve

(1)
| | | | |

(2)
| | | |
| | | | |

(3)
| |
| | | |
| | | | |

(4)
| |
| | |
| | | | |

(5)
| |
| | |
| | | | |

(6)
| |
| | |
| | | | |

Each vertical bar represents a level of metrical strength (or stress). Any of the six alternatives (and of course, many others) could be the grid applied to a particular version of the sentence. The levels of metrical strength of the stressed syllables are categorical in respect of others; that is, for any stressed syllable, there will always be one stressed syllable about which it can be said either that it is less stressed or that it is more stressed than it, but not more/less stressed <u>to a particular degree.</u> This is reflected in the fact that there are no grids which contain steps of more than one (e.g. two or three) levels of metrical strength between a stressed syllable and some other stressed syllable.

For Pierrehumbert (and Beckman), as we have indicated, it is the lexical concomitants of stress which are primary in the generation of prosodic patterns. Prominence, in Pierrehumbert 1980, is a function of metrical strength, as just defined, and emphasis. The latter is a continuously varying parameter, and determines the degree to which a stressed syllable of one level of metrical strength is more prominent than a stressed syllable at the next level down of metrical strength. This is reflected in continuously variable prominence values, such as appeared at the start of the rule sequences in Figs. 3.35-44, and thence directly in continuously variable phonetic values of tonal targets. Pierrehumbert puts it like this: "In an intonation with a H* prenuclear accent and a H* nuclear accent, the nuclear accent could be anywhere from not significantly higher than the prenuclear accent to a great deal higher. What controls this variation is something like 'amount of emphasis' "(1980:39).

Pierrehumbert finds, from analysis of intonation contours, that tonal primitives are sparsely specified compared to the relative prominence specifications of the textual tree. Consequently, they are not organised into a tree, but are sequentially ordered in attachment to the terminal nodes of the textual tree. Thus, they simply act as vectors for the conveyance of underlying prominence (the combination of metrical strength and emphasis). In respect of the discussion of the relative primacy of pitch obtrusion and increases in duration and amplitude, pitch obtrusion, to the extent that it is reflected by the choice and ordering of the tonal primitives and the rules that convert a string of them into an intonation contour, has lowest priority in marking accentuation. In terms of pitch, accentuation is marked merely by the attachment of pitch accents to stressed syllables whose prominence has already been determined. It is in this role that their generation by a finite-state grammar and the generation of the intonation contour by a set of local left-to-right implementation rules is considered appropriate (although, as we have seen, it is the downstep rule which is instrumental in requiring this form of implementation rather than,say, the global application of the elements of a tune stored in an intonational lexicon). The non-local dependencies required in the determination of relative prominences are present only within the textual tree.

That being the case, one might expect that the downstep rule could be done away with, because the relative underlying prominence values appropriate for downstepping sequences could result from (a) metrical strength relations as specified in the textual tree, determined originally, perhaps, on the basis of syntactic considerations, and (b) choices of emphasis, determined originally, perhaps, on the basis of pragmatic considerations. What justification does Pierrehumbert have for incorporating a rule of downstep within the intonational phonology, rather than within a component which determines prominence values?[25]. In this regard, it is appropriate to consider the general appropriateness of the process of downstep in the account of a language such as English.

3.5 THE VALIDITY OF THE CONCEPT OF DOWNSTEP IN ENGLISH

As already noted, Pierrehumbert has incorporated a version of the process of downstep used in analyses of African Tone languages into an account of the intonation of English. However, it is important to note that she hasn't maintained some of the key taxonomic parameters which are used in analysis of particular forms of downdrift or downstep which are specific to individual African languages. Most importantly, she allows English to incorporate sequences exhibiting both total and partial downstep. Figs. 3.16–17 show contours exhibiting total downstep. Here, the second H in a H L H sequence is downstepped 'fully' to the height of the intervening L (whether or not that L is phonetically realised). Figs. 3.18–19 show contours exhibiting partial downstep. Here the second H in a H L H sequence is downstepped only 'partially', as it is higher in pitch than the intervening L.

This usage of the terms 'total' and 'partial' downstep represents a radical departure from the usage found in analyses of African tone languages. In the first place, 'total' and 'partial' downstep are supposed to be mutually

---

[25] Pierrehumbert does present an argument (1980:214, fn. 3) which shows that iterative application of a downstep rule to the underlying prominence value would result in non-local dependencies between prominence values, but this hardly has any force when it is considered that there exist such non-local dependencies anyway on the basis of the metrical textual tree. In any case, the argument is not to the point; what needs to be shown is that downstep is a specifically intonological process, and that it is pivotal in the constraint of the relative height of pitch accents, not just in sequences of downstepping peaks, but in many other sequences which display a local downward step, and is also implicated in the form of rules which can be used to compute all intonation contours.

exclusive possibilities within a language. As Clements says (1980, p.88) : " ... an adequate theory of "total" downstep must account for the fact that all other factors being equal, there is never a systematic pitch difference between the sequences H-L and H-ꜜH or between the sequences L-L and L-ꜜH. It is apparent, therefore, that the pitch interpretation principles governing "total" downstep systems must be different from those governing "partial" downstep systems." (where ꜜ is a diacritic marking downstep on a tone). In the second place, the distance in pitch between a H and following L tone in an African tone language is postulated to be a constant (within a pitch range) for that language. The language could thus in principle be categorised as a 'total' or 'partial' downstep language on the basis of whether the second H in a H L H sequence is at the same level as the intervening L, or somewhere between the pitch of the first H and the intervening L (cf. Clements 1979, p.550).

In Pierrehumbert's analysis of English, such a reference pitch interval between the H and L in a H L sequence in which the L is the second tone of a bitonal pitch accent only exists indirectly in certain contexts: those determined by Rule 10. The downstep constant determines a reference pitch interval for any point within a pitch range that H is fixed at. Her rules also determine another pitch interval, between a H and a following L, when that L is the first tone in a bitonal L+H pitch accent, which could also, in principle, act as some kind of reference interval for determining the quantitative bounds of total downstep. In general, this pitch interval is larger than that found in a H L sequence where the L is the second tone of a bitonal pitch accent; the L tone in the former is close to the baseline, whereas that in the latter isn't. Yet is in that context (H L+H) that the process of partial downstep is deemed to occur, whereas in the other context (H+L H), total downstep is said to occur.

This apparently confusing state of affairs could be emended by a different choice of terminology, or simply by dropping the requirement that a division between different types of downstep occurs in English. Then, analysis could concentrate on corroborating the claim that there is a constant factor that governs the pitch heights of consecutive H accents in a sequence of alternating H and L accents. However, that would be to ignore the fact that there appear to be qualitative , as well as quantitative, differences in the L

tones in H L+H sequences, as compared with those in H+L H sequences, under Pierrehumbert's analysis of English.

In H+L H sequences, the height of the second H tone and the intervening L tone (assuming it is manifested) are determined by the same downstep constant, as we have seen. In the absence of any differences in prominence between the two pitch accents, they are generated with the same pitch height, and a sequence of such sequences would result in a train of asymptotically declining L tones as well as H tones. In H L+H sequences, the height of the second H tone and the intervening L tone are determined by a different constant n (< the downstep constant k). This ensures that, in the absence of any differences in prominence between consecutive pitch accents, a sequence of L+H accents results in a train of asymptotically declining L tones, with a shallower slope than that of the train of L tones in a sequence of H+L H sequences over the same text. Fig. 3.27 summarises the situation:

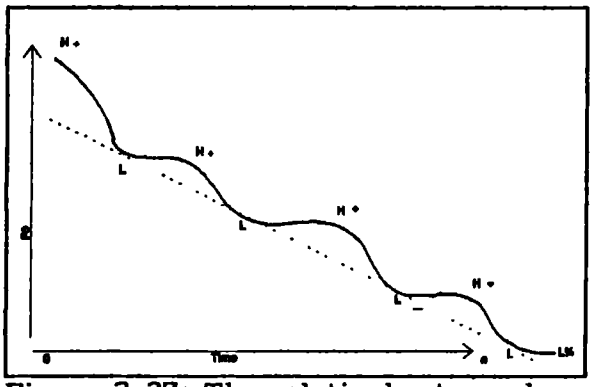


Figure 3.27a The relatively steep slope through the Ls in a sequence of H+L pitch accents (where the pitch accent is H+L*).

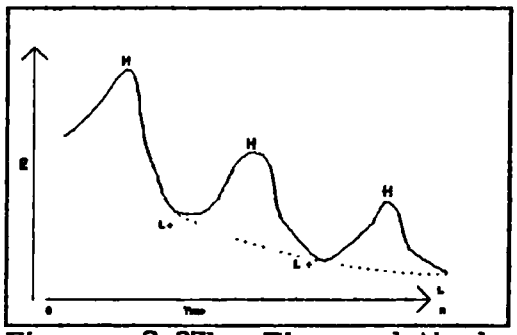


Figure 3.27b The relatively shallow slope through the Ls in a sequence of L+H pitch accents.

When there are differences in prominence in successive pitch accents in the sequence, the qualitative difference in the behaviour of the L tones shows up. If, for instance, the second accent in each of the contours in Fig.3.27 has an underlying prominence value less than that of the other two, the patterns in Fig. 3.28 result.

The fact that in one case the curve passing through the L tones is convex relative to the bottom of the graph and in the other it is concave is a function

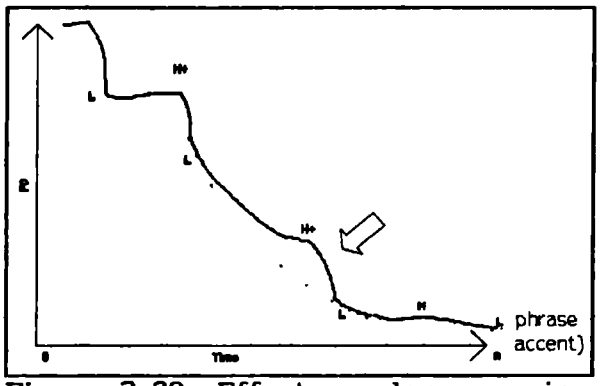


Figure 3.28a Effect on slope passing through Ls of reduction in prominence on a medial pitch accent (arrowed). (H+L sequence, pitch accent is H+L*)

of the fact that the L tones in the second case exhibit behaviour consistent with that found in monotonal L pitch accents (the more prominence, the lower in pitch, the less prominence, the higher in pitch they become), whereas in the first they exhibit behaviour consistent with that found in monotonal H pitch accents (the more prominence, the higher in pitch, the less prominence, the lower in pitch they become)[26].

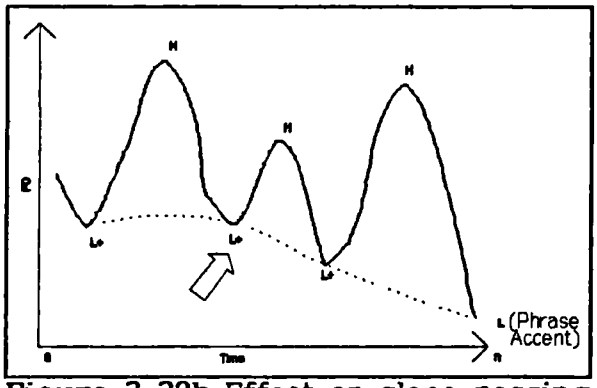


Figure 3.28b Effect on slope passing through Ls of reduction in prominence on medial pitch accent (L+H sequence)

There is another difference between these types of sequence resulting from prominence differences, and that is (it seems to the current author) that the former are much less likely to occur than the latter. The reason for this is that, putting aside the problems of instability inherent in Rule 4 (which is involved in generating sequences of the second type) which could in any case be solved by an appropriate adjustment, the former is more unstable than the latter. This is because, if a reduction in prominence is carried through onto subsequent pitch accents, or if, in the worst case, there is an increasing reduction in prominence on consecutive pitch accents in a sequence of H+L H sequences, the phonetic values of the H tones would become increasingly diminished to the point that they would be in danger of adopting values only appropriate for L tones (there would be many tones in a prolonged sequence very near the declining baseline). This is not so in the case of the second type of sequence; if a reduction in prominence is carried through onto subsequent pitch accents, the line

[26] For a critique of this dual interpretation of L tones within Pierrehumbert's framework, and a reanalysis which dispenses with it, see Grice 1992 Ch. 6.

passing through the L tones naturally ends up higher than it would have done had the prominence on consecutive pitch accents been identical. Also (assuming an appropriate adjustment in Rule 4), an increasing reduction in consecutive pitch accents results in the H and L tones both tending towards a medial asymptote, which is a perfectly stable situation. What's more, that kind of behaviour, resulting, in this model, from variable prominence assignments to consecutive accented syllables, is much more likely to occur with sequences of H L+H sequences than with H+L H sequences. Thus, the sorts of contours in Fig. 3.29 could all result from a sequence of H L+H sequences:

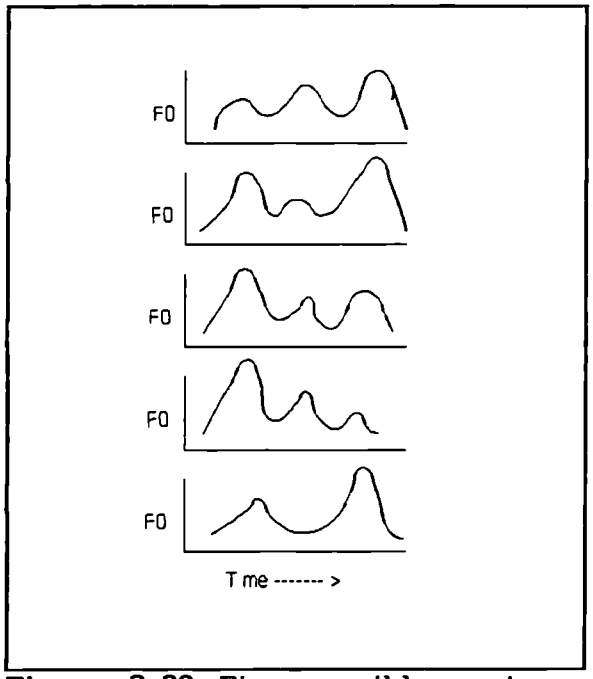


Figure 3.29 Five possible contours consisting of a sequence of L+H pitch accents.

The H L+H sequence thus has far fewer constraints on the heights of consecutive accents than the H+L H sequence does. In fact, there is no reason to suggest that the so-called downstepping sequence involving H L+H configurations is any more common than any other type of sequence involving the same.

With the H+L H sequence, a downstepping sequence with an asymptotic decline of the constituent tones is the only generally observed pattern. The only real variant, given the inherent instability of the sequence once iterative reductions in prominence are applied, is a pattern in which there is increasing prominence from lower to higher, as in Fig. 3.30:

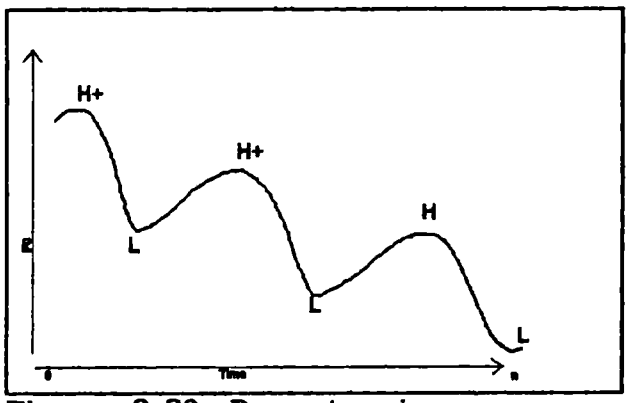
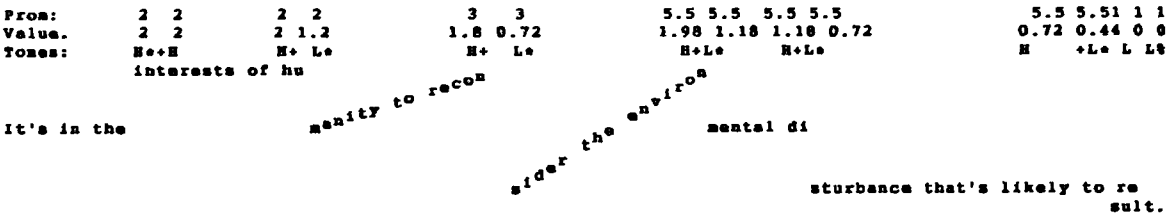
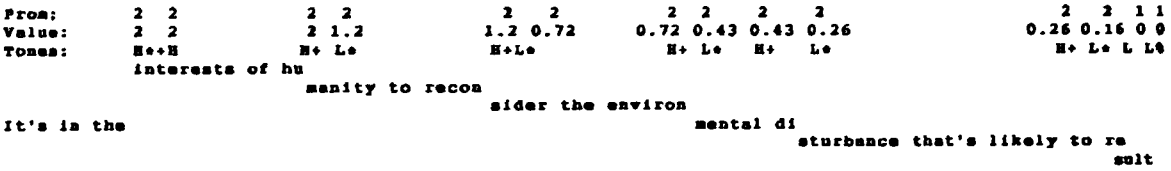


Figure 3.30 Downstepping sequence (comprising H+L* accents) underpinned by a prominence pattern increasing from lower to higher.

Even here, though, there are constraints on the form of sequences that are allowed. In particular, consecutive H tones don't tend to disrupt a downward moving trend line, in the way that is done in the following:



Unless, perhaps, it were being spoken with constant side interruptions, the above sentence would sound as though it were spoken by a person who was a little unhinged, whereas the following is more natural:



An additional problem here is the conflation of the effect on F0 values of underlying prominence values with that of the operation of the phonological and phonetic rules, which Pierrehumbert freely acknowledges:

"Downstepped tones are themselves subject to expressive variation in level. Thus one of the problems which English presents is how to describe the interaction of relative prominence and downstep in controlling tonal value." (1980, p.150).

This is a problem which Silverman addresses (Silverman 1987) and Ladd describes (1990), and which will not be discussed in detail here. It is, however, a general problem that afflicts the process of the interpretation of any of the contours generated by Pierrehumbert's model.

In fact, given the observations that have just been made, there is a solution to the problem of the conflation of prominence and downstep. It has already been shown that the patterns of relative peak height in sequences of H L+H sequences include a downstepping sequence as one amongst many surface options, whereas this is not true of sequences of H+L H sequences. It is not unreasonable to suggest, therefore, that the downstep rule should be restricted in its applicability to the latter type of sequence. In that case, many of the cases of conflation of downstep with prominence would disappear; it has already been seen that to do away with the downstep rule is effectively to do away with Pierrehumbert's whole framework, and so the relative heights of accented syllables in H L+H sequences could be computed globally or locally without reference to predecessors (this latter option would require some new constraints to replace the downstep rule, though). This would leave downstep applicable only in the case of H+L H sequences.

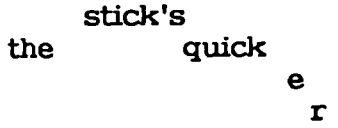
In the above consideration of constraints on the form of downstepping sequences, the H+L* H sequence was concentrated upon. The H*+L H sequence is more problematic when it comes to identifying it as part of a downstepping sequence, and so discussion of it has been postponed until now. The reason it is difficult to identify is that there is no necessary pitch obtrusion at the points of the accented syllables in a sequence of H*+L pitch accents, other than at the beginning and end. Thus, the sequences

H*+L H*+L H*+L H*+L H*+L L L%

and

H* L* L L%

could have equivalent surface forms. Furthermore, because the L tone in the sequence H*+L H is not physically manifested in the contour, there are many cases of ambiguity in tonal sequences of that form. For instance, it is not clear whether the contour



should be represented by the sequence

H*+L H* L L%

in which downstep is operative, or the sequence

H* H* L L%

in which the scaling of the accents reflects directly the underlying prominence.

These alternatives and ambiguities throw up the problem of justifying the existence of the H*+L tone. Additional problems are the obvious one that it is the only pitch accent in which the L tone has no phonetic realisation, and the fact that something like constraint 6, mentioned in section 3.2 above, is required to account for the interpolation between H* and the following H occurring in the requisite way (skipping the non-appearing L), even though the form of constraint 6 is bound to be more complicated than it is presented, because in its present form it would preclude the assignment of phonetic value to isolated tones not matching the SD of any of the rules, such as appear in the Surprise-Redundancy contour.

Justification for its existence would certainly be found if it could be argued to occur in other downstep contexts. Yet in all of Pierrehumbert's figures, the only case in which it so appears is her Fig. 4.33, reproduced here as Fig.3.31. However, there is no reason to suppose that the initial pitch-accent should not be a H* monotonal one in this case, particularly if it is accepted that L*+H pitch accents are freed from being necessarily implicated in downstep contexts, such as would arise from the resultant H* L*+H sequence (and even if they weren't so freed, the contour could still be presented as a case of downstep resulting from a H* L+H sequence with a high underlying prominence value causing the high F0 peak on the word "undergraduate".

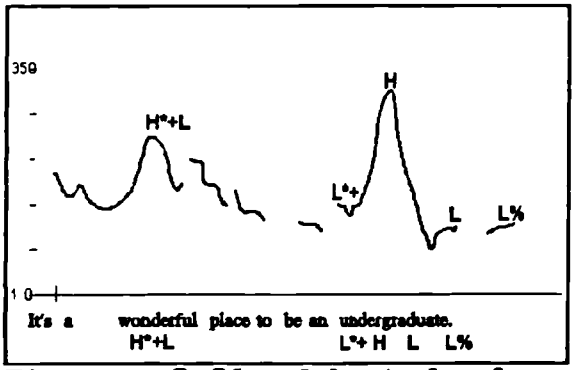


Figure 3.31 Adapted from Pierrehumbert 1980, fig.4.33, p.351.

Thus, it is difficult to find justification for the existence of the H*+L accent. If the L+H accents are removed from the context for the downstep rule, even Pierrehumbert's original rather weak reason for its inclusion into the inventory of pitch accents, to "fit the only remaining peg into the only remaining hole" (1980, p.152) loses some of what force it has. For these reasons, it is proposed that the H*+L could be dispensed with. This just leaves H+L* H sequences as the context for downstep. This sequence itself is not without its problems. As we have seen, there are constraints on the kinds of prominence variation that can underlie sequences of such pitch accents. In fact, in its canonical form (as a stepping head), it can be seen that the prominence values assigned to all the pitch accents must be the same (as in Fig. 3.16), otherwise the asympotically declining contour is not generated. However, if that is the case, then a further, fundamental motivation disappears of (a) the downstep rule, and (b) underlying prominence values not directly determined from F0 values. For Pierrehumbert argues that in downstepping sequences, the existence of the downstep rule and the possiblity of varying underlying prominence independently of F0 values in pitch accents explains why the last accented syllable (the nucleus) is the most prominent, even though it is the lowest in pitch, and that is because increased underlying prominence on the nucleus relative to the other accents (this prominence relationship presumably applying all the way back to the initial accent) interacts with downstep to produce the requisite prominence-pitch ratio. But the nucleus can't be the most prominent when it is the lowest in pitch because, to produce that stepping head and nucleus pattern, equal prominence values are required. And they are required not just because, as it turns out, equal prominence values reflect the rather stylised nature of such sequences in which no information is imparted, but also by Pierrehumbert's very rule set.

There remains very little motivation for maintaining the existence of underlying prominence values independent of F0 values. In answer to the question posed at the end of the previous section, then, Pierrehumbert might as well apply the downstep rule to prominence values as to F0 values, in the contexts in which it is applicable. If it is applied to prominence values (as specified in the metrical tree determining metrical strength relationships and by additional emphasis) then there would be a direct mapping from prominence values to F0 values, with no rules in the intonational phonology.

If it is applied to intermediate F0 values, then there would be heavy constraints on the permissible prominence values generated from the metrical tree (generally, the prominence values in a downstepping sequence would all have to be equal) and on those emphatic factors of prominence determined pragmatically (they would have to be such as not to disrupt the asymptotically declining downstep pattern).

3.6 CONCLUSION

In this discussion, it has been seen that downstep is a valid concept in the analysis of English intonation, but only in restricted contexts, and probably not in the pervasive way that Pierrehumbert has suggested. In particular, downstepping sequences most often occur in stepping heads, though they can occur in sequences which Pierrehumbert would analyse as involving the sequence of tones H L+H. In Chapter 5, the possibility of the existence of isolated instances of downstep (i.e. those not part of a downstep sequence) is examined during the development of a performance model of local accentuation and declination; there, separate abstract prominence values play no part, and the problem of the conflation of prominence values with the operation of downstep, alluded to towards the end of this chapter, and a problem specifically for a competence model of intonation, is circumvented.

CHAPTER 4
LOCAL VS.GLOBAL DECLINATION

4.1 INTRODUCTION

Chapter 2 referred to the declination effect, whereby an accented syllable later in an utterance is considered more prominent than an earlier one if it has the same peak pitch (and assuming the same basic shape of accent). That effect obtains in both production and perception; a speaker will naturally reduce the peak pitch of a second accented syllable compared with a first in order to impart the same measure of prominence on the two syllables, and a hearer will expect such a reduction.

In this chapter, the declination effect will be examined predominantly from the hearer's point of view. In particular, the third issue raised at the end of Chapter 2 - to what extent a global declination function is derivable from the interaction of local functions - will begin to be examined by a study of relevant literature in the topic and by means of a perceptual experiment. At the end of the chapter, some of the issues raised will be examined in respect of both production and perception, in preparation for the introduction in the next chapter of a model of prominence-pitch relationships which takes account of both processes.

4.2 PERCEPTUAL INVESTIGATION OF THE EFFECT OF VARYING DECLINATION SLOPES

Certain researchers have already performed experiments which go some way to investigating the effects of variation in specifically local declination slopes on the declination effect. Their aim was not to investigate the relationship between local and global phenomena, their experiments always manipulating global declination slopes. However, in so doing, they thereby manipulated the local slopes around accent peaks, and it is because of this that their results are relevant.

4.2.1 Leroy (1984)

In the two experiments reported on, Leroy (1984) investigated the hypothesis that declination was a 'psychologically real' phenomenon, rather than a perceptually requisite operative construct, as the founders of the Eindhoven school had originally intended (see Cohen et al. 1982). In the first

experiment, she asked the question that Pierrehumbert had posed five years previously (Pierrehumbert 1979), viz. whether listeners compensate for the declination effect when judging the relative pitch height of successive accented syllables in an utterance. However, she made a more detailed examination of this question by testing specific hypotheses about how listeners would make such judgments, both in the presence of declination slopes of differing magnitude, including zero, and also without any reference slope at all against which to frame the perceived pitch of the accents. Her basic hypothesis was that for the declination effect to take place, there has to be an observable decline in the F0 contour in a low part of the speaker's range - a physically present declining baseline - and that the more the baseline declines, the stronger will the declination effect be (that is, the less high will the second accent need to be relative to the first for them to be judged equally high in pitch).

The stimuli she used were LPC diphone-synthesized reiterant speech tokens comprising seven 'ma' syllables, with the second and penultimate syllables accented with 'pointed-hat' accents.

In the first condition (her condition 'N'), stimuli were prepared with a 'normal' declination slope (that is, according to the specification of the Eindhoven school as stated in Chapter 2) of -3.35 semitones per second. There were seven stimuli, in which the peak fundamental frequency on the first accent was 140Hz and that on the second was varied in 1.56 semitone steps, four below and two above the 'objectively equal' stimulus value of 140Hz. The baseline began at 108 Hz and ended at 75 Hz. The degree of 'drop' on the baseline between the accent peaks - that is, the drop between the projection from one peak onto the baseline and the projection from the next onto the baseline (the 'interaccentual peak baseline drop') was 3.21 semitones. It is estimated from Leroy's figures that the 'local' declination slope between the bases of the accents (the 'local interaccentual baseline drop') was 1.25 semitones.

In the second condition (her condition 'S'), stimuli were prepared with a 'steep' declination slope of -5.78 semitones per second. There were six stimuli, with the first accent having peak fundamental frequency of 172 Hz, and the second having a value which varied in the following semitone steps

(from low pitch to high), four below and one above the 'objectively equal' stimulus value of 172 Hz : -8.22, -5.39, -2.72, -1.36, 0 (= 172Hz), 2.61. The baseline began at 140Hz and ended at 75 Hz. The interaccentual peak baseline drop was 5.78 semitones. The local interaccentual baseline drop is estimated to have been 2.15 semitones.

In the third condition (her condition 'M'), stimuli were prepared with a monotonous baseline, i.e. with no declination slope. There were six stimuli, with peak fundamental frequency on the first accent being 128 Hz, that on the second varying in 1.56 semitone steps, three below and two above the 'objectively equal' stimulus value of 128 Hz. The baseline was set at 90Hz. The interaccentual peak and local interaccentual baseline drops were thus both 0 semitones.

In the fourth condition (her condition 'IS'), the pointed-hat accents were isolated from the surrounding declining baseline and presented with just a silent period between them of the same duration as occurred between the accents in condition 'N'. The stimuli were prepared by excision from the stimuli in condition 'N', and thus numbered seven, with the same F0 values on the first and second accent peaks.

The first three conditions were tested separately from the fourth. Three instances of each stimulus in conditions 'N', 'S' and 'M' (= 3x7 + 3x6 + 3x6 = 57), and five of each in condition 'IS' (= 5x7 = 35) were thus separately randomized and recorded and presented to 28 student volunteer subjects. Their task was to say, for each stimulus, whether they judged the second accent to be higher, less than or equal in pitch to the first.

From their judgments, a psychometric labelling curve was determined for each of the three possible judgments in each condition. An example of the curve for condition 'N', pooled over the data of all 28 subjects, appears in Fig. 4.1 (adapted from Leroy's Figure 3; on the y-axis, maximum N corresponds to subjects' x repetitions = 28x3 = 84).

In order to extract data in a form appropriate for a unitary labelling curve, from which a Point of Subjective Equality (PSE) for each condition could be determined, which identifies the point on the x-axis at which the objective

pitch difference between the accent peaks (in semitones) corresponds to perceived equal pitch on them, the 'equal' scores had to be apportioned appropriately to the 'higher' and 'lower' scores respectively. Leroy chose to do this after she had translated the pooled raw data into z-scores, that is, points on a standard normal distribution curve. The method of apportionment was to take as a final z-score value the mean of the z-score for the 'higher' judgments and that for the combined 'higher' and 'equal' judgments, effectively weighting the 'higher' judgments by a value 2 and the 'equal' judgments by a value 1.

After weighting the z-score data in this way, a linear regression was performed on them, and the PSE identified as the point at which the regression line passed through the zero-point on the y-axis, that is, the centre of the standard normal distribution, which corresponds to the 50% point on the 'higher responses

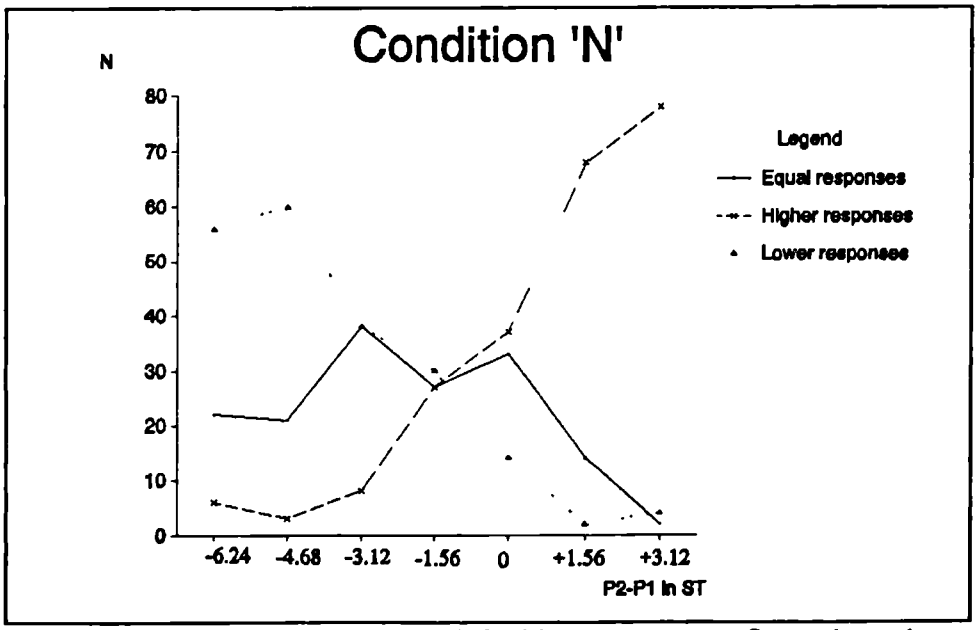


Figure 4.1 - Psychometric labelling curves for the three response types in the Normal declination condition of Experiment 1 in Leroy (1984).

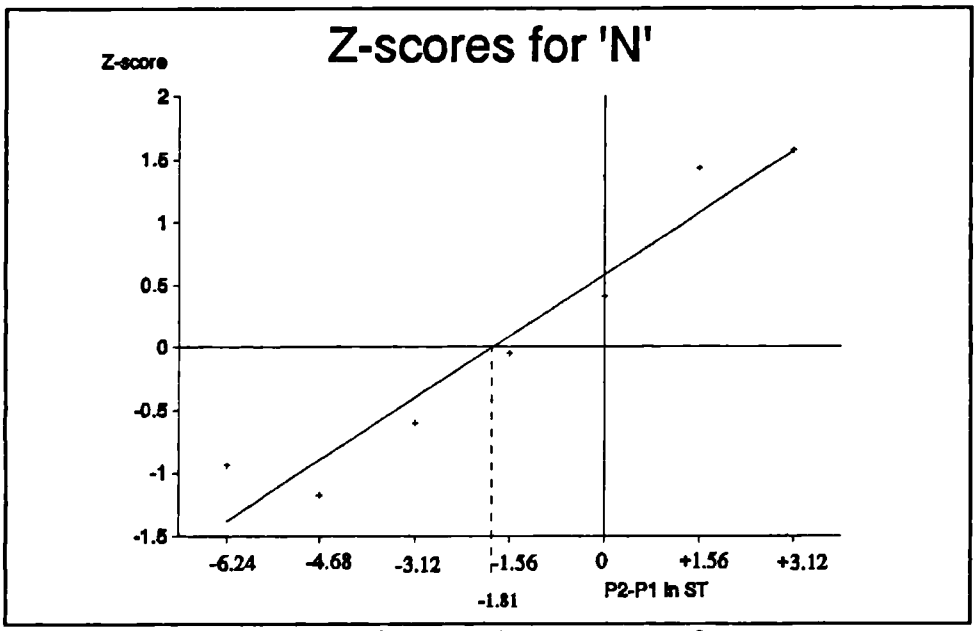


Figure 4.2 - Z-scores for Condition 'N' of Experiment 1 in Leroy (1984), and linear regression line fitted to them (adapted from Leroy's Figure 4).

judgment' axis[1]. Figure 4.2 depicts the regression line fitted through the weighted z-score data for condition 'N' (adapted from Leroy's Figure 4). The PSE is identified as -1.81 ST, with a standard error of measurement of 0.89 ST.

The equivalently determined values of all four conditions were as follows :

CONDITION	PSE	s.e.m.
S	-2.08 ST	0.55 ST
N	-1.81 ST	0.89 ST
M	-1.03 ST	0.41 ST
IS	-0.32 ST	0.58 ST

Leroy applied t-tests to the six pair-wise permutations of these conditions, and found that only the difference in the result between conditions S and IS could be considered significant ($t=2.2$, $df=11$, $p < 0.025$). The differences between M and S ($t=1.53$, $df=11$) and N and IS ($t=1.4$, $df=12$) were nearly

[1] This procedure is equivalent to that of fitting a cumulative normal distribution curve to the raw response data, which was done in experiment 1 in Chapter 2.

significant (.1 > p > .05). When the data for conditions S and N were pooled, and likewise those for conditions M and IS, the results were as follows :

CONDITION	PSE	s.e.m.
(S+N)	-1.97 ST	0.48 ST
(M+IS)	-0.5 ST	0.68 ST

These pooled conditions were found to yield significantly different results at the p = 0.05 level (t=1.76, df=24)[2].

4.2.1.1 Leroy's conclusion

Leroy takes these results as indicating that the declination effect is partly mediated by the presence of a declining baseline between accents (whence the difference between those cases where there is declination - conditions S and N - and those where there isn't - M and IS). She also concludes, on the strength of the less-than-zero PSE for the latter two conditions, that the effect is partly attributable to the influence of an expected but not perceived, and thus abstract, declining reference baseline that listeners have internalised. Finally, she notes that the declination effect is not so strong as would be predicted by the amount of declination in the stimuli; the interaccentual peak baseline drop for condition N is 3.21 ST, yet condition N yields a declination effect of only 1.81 ST, and the interaccentual peak baseline drop for condition S is 5.54 ST, yet condition S yields a declination effect of only 2.08 ST[3].

4.2.2 Gussenhoven and Rietveld (1988)

A contrasting result - that the declination effect is stronger than predicted by the amount of declination in stimuli tested - is found in the Gussenhoven and Rietveld's 1988 experiment. The three hypotheses the experiment was designed to test were (i) that the declination effect is due to time-dependent declination, such that two-accent utterances with a longer interaccentual

---

[2] There is reason to question the accuracy of Leroy's results. In Appendix 4.A appears a reanalysis of her data which shows that her results are less significant than she presents them to be.

[3] At this point it is worth remembering that the local interaccentual baseline drops for conditions N and S are, respectively, 1.25 ST and 2.15 ST.

stretch show more of a declination effect than those with a shorter one; (ii) that the declination effect is due to a peak-by-peak lowering function on non-initial peaks (similar to downstep, but of smaller magnitude), and (iii) that the declination effect is due to final lowering of the last 500ms or so of the contour.

The main perceptual experiment in their paper used reiterant speech stimuli of the same basic form as Leroy's, but with one less or one more interaccentual 'ma' syllable (so the stimuli were of the form 'ma MA ma ma MA ma' or 'ma MA ma ma ma ma MA ma'). Also, they were not prepared by diphone synthesis, but by analysis and resynthesis of a selected token in a prior production experiment. Two different parameters were chosen to elicit results bearing on the three hypotheses: the duration of the interaccentual stretch (as has just been indicated), and the peak height of the first accent in the utterance. These parameters both had two values, as indicated in the schematized representation of the stimuli in Figure 4.3.

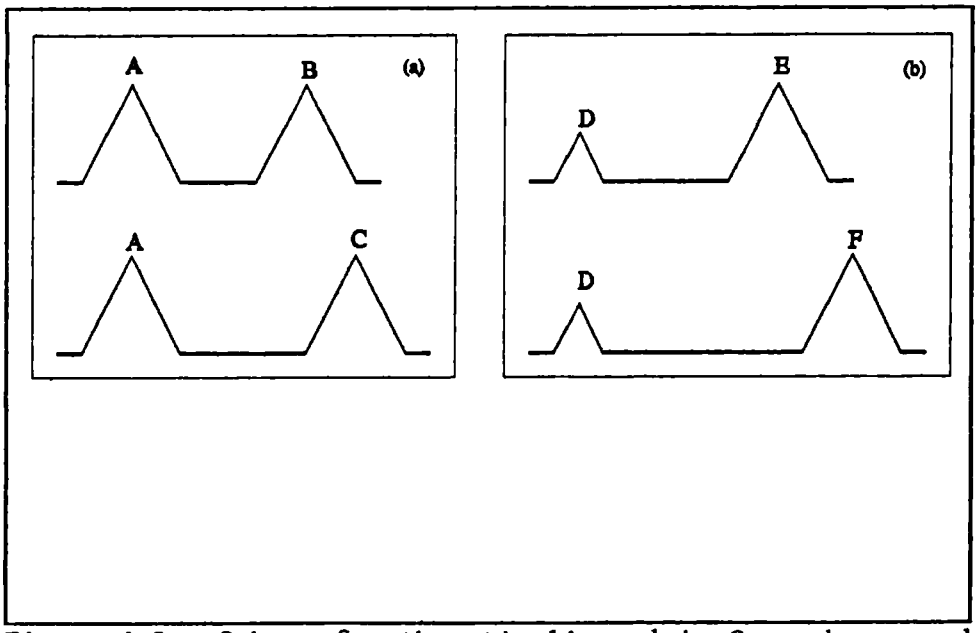


Figure 4.3 - Schema for the stimuli used in Gussenhoven and Rietveld's (1988) main perceptual experiment (adapted from their Figure 1). In (a) A=B=C in F0 value. In (b), E=F likewise.

The local interaccentual baseline was either short (two syllables = c.250ms) or long (four syllables = c.500ms). (The distance between the peaks was 500

ms in the former case and 750 ms in the latter case.) The first accent peak was either equal in height to the second or lower than it.

The way the values of these parameters were designed to test the plausibility of the three hypotheses was as follows. If the declination effect was a function of time-dependent declination, then the prominence rating elicited by the second accent in the long utterances (C and F) would be proportionately greater than that elicited in the short utterance (B and E), and all second accents would elicit greater prominence ratings than the accent A. If the declination effect was a function of a peak-by-peak lowering effect, then the prominence rating elicited by the second accent in the equal peaks utterances (B and C) would be proportionately greater than that elicited in the unequal peaks utterances (E and F). And if the declination effect was the function of a final lowering effect, then the prominence rating elicited by the second accent in all utterances would be proportionately greater than that elicited by the A accents, but equal amongst themselves.

The prior production experiment was used to determine appropriate values for different pitch range values for the respective accent peaks. Details of the precise durational parameters of the resynthesized contours used in the perception experiment are not all available in the paper, but the canonical form of the basic stimulus can be constructed using informed guesses about final lengthening, given that the total duration of the short utterance was 1120ms and that of the long 1370 ms. This canonical form, with salient measurements, appears in Figure 4.4.

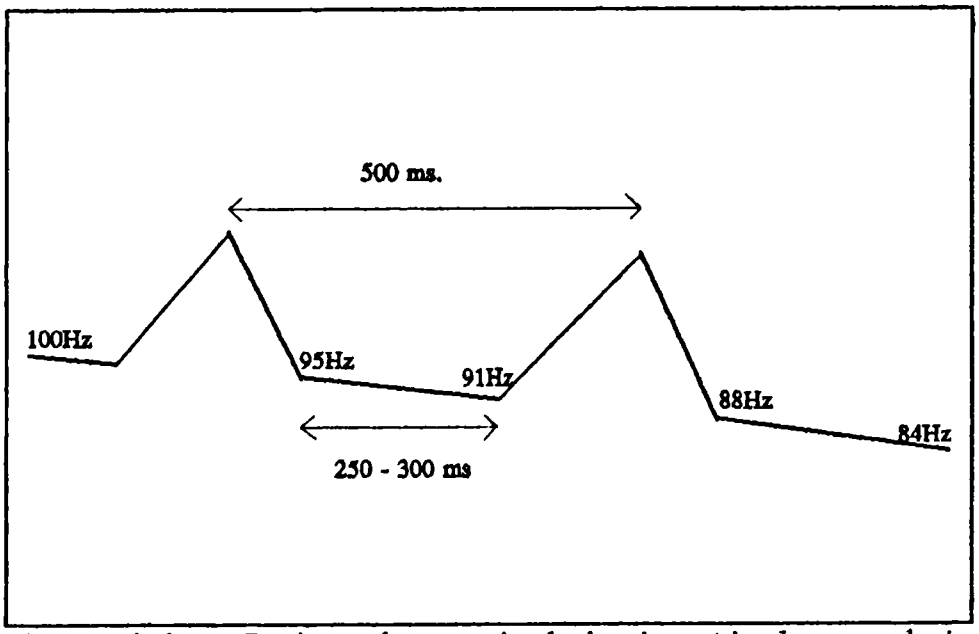


Figure 4.4 - Estimated canonical basic stimulus used in Gussenhoven and Rietveld's 1988 perceptual experiment.

The long utterance was created from this short form by splicing in a repetition of the two interaccentual syllables. The amount of drop in the slightly declining baseline used in this experiment was not adjusted for the increase in length. Thus the local interaccentual drop was of the order of only 4Hz (0.74 ST) in both the short and the long utterances, though the slope becomes shallower in the longer utterances.

The short and long utterances were each resynthesized in nine pitch ranges, once for the equal peaks condition and once for the unequal peaks condition. The accent peak values were the same for short and long utterances, and were as follows (as adapted from Gussenhoven and Rietveld's Table II) :

	"Equal-peaks"		"Unequal peaks"	
No.	First	Second	First	Second
1	182	175	158	175
2	177	170	154	170
3	172	165	150	165
4	166	160	146	160
5	161	155	142	155
6	156	150	138	150
7	150	145	134	145
8	145	140	130	140
9	140	135	125	135

In the equal-peaks condition, prominence judgments were required of both accent peaks, but only the second needed judging in the unequal-peaks condition. As each elicitation only warranted one prominence judgment (so as not to tax the subject), there were two repetitions of 2x9=18 stimuli in the former condition and one in the latter, making 3x18 = 54 stimuli in all. These were all repeated three times, so that the total number of stimuli presented randomly to an effective group of 17 student listening subjects was 54x3=162. Their task was to rate an indicated accent peak in each stimulus for prominence on a 10-point scale.

Results were pooled across subjects, and across conditions in the following post hoc comparisons (values in brackets are mean prominence ratings elicited for the indicated accent peaks):

(a) First peak vs. second peak - A + A (5.85) vs. B + C (6.23) $p<0.01$

(b) Equal peaks vs. unequal peaks - B + C (6.23) vs. E + F (6.08) n.s.

(c) Short contour vs.long contour - B + E (6.05) vs. C + F (6.26) $p < 0.05$

From these results, Gussenhoven and Rietveld conclude that the declination effect as observed by Pierrehumbert 1979 has occurred in the data (row (a)), that it is not attributable to a peak-by-peak lowering effect (row (b)), and that it is at least partly attributable to a time-dependent declination function (row(c)). On the basis of the difference between the mean prominence rating for the early second peaks and that for the late second peaks (6.05 - 6.26 = c. 0.2 scale points difference corresponding to a difference in duration of 250ms, i.e. 0.8 scale-points/sec.), and the scale-point:Hz. ratio estimated from their results (= 1:12),they compute a time-dependent declination effect of between 9 and 10Hz per second. Given that the distance between the peaks A and B in the short contour is 500 ms, the difference between peaks A and B eliciting equal prominence ratings if the declination effect were <u>wholly</u> attributable to a time-dependent declination function should be between 4 and 5 Hz. Instead it is about 10Hz. On the basis of their original hypotheses, they conclude that the remaining 5 or 6Hz is attributable to a final lowering function.

This conclusion has been criticised (Terken 1989a) on the grounds that the use of the pooled equal peaks and unequal peaks data for computing mean prominence ratings for the short and long contours is unwarranted, because of the spread of the two data values used to compute the mean in the long

contours case (6.39 vs. 6.12). Terken also shows that if the equal peaks data and unequal peaks data are taken separately, linear regression of peak fundamental frequency on prominence rating from the published data yields a difference between first and early second peak, for the same prominence rating, of 11-14Hz (depending on where in the range you are). In addition, the data yield plots of peak F0 on first peak, early second peak and late second peak through which, at whatever prominence level, straight lines can be drawn which decline at a rate, for instance, of 20Hz/s at prominence level 5 and 15Hz/s at prominence level 7. This argues, he says, for a unary account in terms of time-dependent declination for the equal-peaks contours and lack of declination for the unequal-peaks contours. Final lowering doesn't need to come into the picture.

Gussenhoven and Rietveld's reply (1989) points out that the pooling of the equal peaks and unequal peaks mean prominence ratings is justified since the difference between the two prior to pooling was not found to be significant in the initial ANOVA. Further, they claim that their aim has been to test for a single theory of declination, appropriate for all utterances, which Terken's division into an account for equally accented contours and one for unequally accented contours militates against.

In fact, notwithstanding statistical considerations, Gussenhoven and Rietveld may have been too hasty in attributing some of the declination effect they identified to Final Lowering. For there is nothing to prevent that part of their identified declination effect being attributed to an abstract internalised declination function, in the same way as was done by Leroy. For them to have been able to attribute it to final lowering, they should have shown that the degree of the final lowering effect was the same as that physically present in the contour. But how is final lowering to be measured from the contour? The canonical form in Figure 4.4 shows no two-piece declining function the second part of which is identifiable as a final lowering 'trap-door'. One must assume that the intention was that the final lowering was interpreted as having been abstractly compensated for by subjects.

Further consideration shows that neither Gussenhoven and Rietveld nor Terken, in their discussion of this experiment, have addressed the crucial issue which Leroy took into consideration, viz. the question of what in the

declination effect is attributable to pitch phenomena physically present in the F0 contour, and what is compensation according to an abstractly computed reference declination function. At least Gussenhoven and Rietveld attributed some part of the declination effect to physically present declination, even though they then attributed the rest to an unspecified function which they wished to identify as final lowering. Terken attributes the whole of the declination effect to a declination function which has no physical counterpart in the F0 contour, which is, as we have noted in Chapter 3, a modelling exercise which is characteristic of the Eindhoven school. However, it ought not to be applied to new experimental data without further elucidation.

At this stage, an interesting observation is that in Gussenhoven and Rietveld's data, the physically present declination only partly accounts for the declination effect, whereas in Leroy's data, the converse is found. The amount of declination suggested by the observed declination effect is only part of the physically present declination, although it should be remembered that the slope of Gussenhoven and Rietveld's declining baselines (-2.68 ST/sec for the short utterances and -2.19 ST/sec for the long utterances) were shallower than those of Leroy (-3.35 ST/sec for condition N and -5.78 ST/sec for condition S).

4.2.3 Terken (1989b,1991)

In the two papers cited, Terken reports on a set of experiments which asked more general questions about the relative prominence of two adjacent accents, but which included consideration of the declination effect. He asked whether the relative prominence of an accent is due to (a) the pitch it reaches on the accent peak, compared with that of its neighbour, or (b) the distance between the accent peak and the baseline. At the same time, he asked

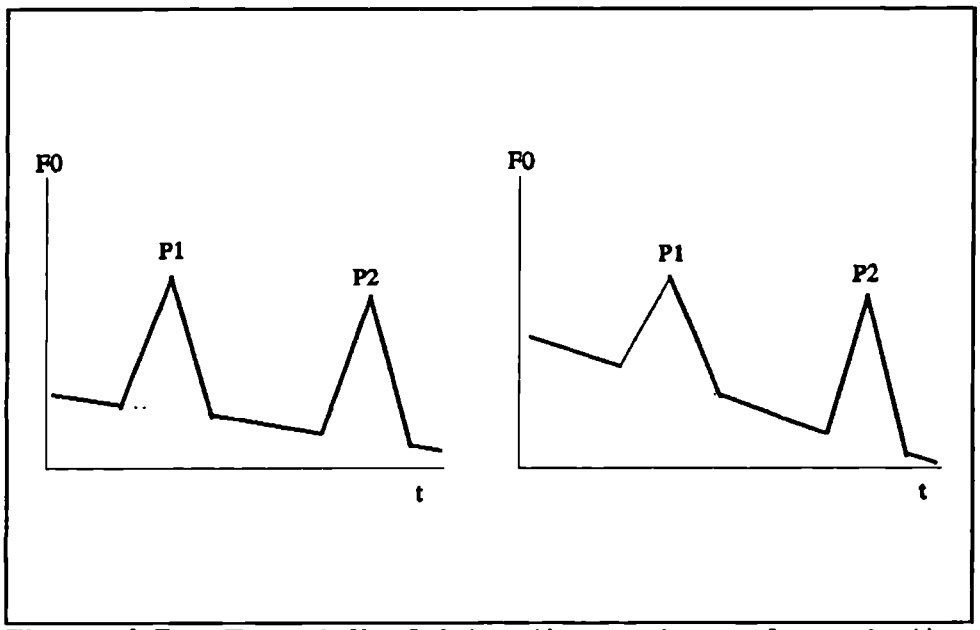


Figure 4.5 - Two stylised intonation contours demonstrating purpose of the experiments in Terken (1989, 1991). Both P1 and P2 have the same absolute values, but the heights of the respective peaks from the baseline differ.

what effect variation in the slope of the baseline would have on the relative prominence of the accents. The importance of the effect of that variable on the other two can be seen in the two example F0 contours which appear in his figure 1 (in both papers), here redrawn in Figure 4.5. In both contours, the difference in peak pitch on accents P1 and P2 is the same; however, in the second contour, the steeper baseline means that the distance from the peak of P2 to the baseline is greater than that from the peak of P1 to the baseline, whereas in the first, the two distances are the same. If the relative prominence of P1 and P2 depended only
on peak height, then P2 should be given the same prominence judgment relative to P1 in both contours. If, on the other hand, it depended only on the relative distance between the peaks and the baseline, then P2 should be given a higher prominence judgment relative to P1 in the case of the second contour.

The way he set about eliciting appropriate judgments in respect of these two hypotheses (which, in Terken 1991, he refers to as the MAX and CHANGE hypotheses respectively) was to conduct two pairs of tests in which the subject had to match a second accent to a first. In the first pair, there was a flat baseline, and there were eleven different pitch heights of the P1 accent

peak to which the subject had to match P2.[4] In the second pair, the same eleven peak values on P1 were tested, but the slope and starting value of the baseline were varied so that the distance between P1 and the baseline was more or less the same in all eleven conditions. In Figure 4.6[5], the contours in the two different pairs of tests are schematised, and Table 4.1 lists the values of the salient points in each of the test pairs[6]. Terken also addressed the question of the difference between pitch and prominence judgments: in the first test of each pair, the subject was asked to match P1 to P2 in pitch, and in the second to match P1 to P2 in prominence. In Table 4.1 the P2 values are estimates of the mean values elicited in the prominence-matching task of the second pair of experiments.

---

[4] The matching was done by selection from a pool of nineteen resynthesized tokens, stored on computer disk, in which the peak F0 on P2 varied, in steps of a few Hz., below and above the P1 value.

[5] It is not easy to represent the compromise between approximately equal linear intervals between values for P1 and approximately equal logarithmic intervals between P1 and B1, referred to below in footnote 6, in the lower part of Figure 4.6, which has been devised by the author. The equal linear intervals between values for P1 have been emphasized at the expense of apparently increasing P1-B1 intervals.

[6] The values of P1 in the two sets of eleven different stimuli were chosen to be as equally separated as possible on a linear Hz scale; that is, as far as the sampling rate of 10Kz allowed (although it is not clear why the third-from-top value could not have been set at 147, making the minimum spacing between P1 values 6Hz rather than 7Hz, whilst maintaining equal D1 values in semitones, as mentioned next). The values of D1 in the eleven stimuli were chosen to be the same, as far as possible, on a semitone scale (rounded down to the nearest unit, all values of D1 are 6 semitones). The reason for the difference in the scaling of the salient values appears to have been that Terken considers the hypothesis that relative prominence is a function of differences in peak pitch to be proposed as valid specifically in the Hz domain, whilst the hypothesis that it is a function of differences in the distance from peak pitch to the baseline to be proposed as valid specifically in a logarithmic domain - "because musical intervals are equal on a logarithmic scale" (Terken 1991, p.1772). Similarly, the slopes of the varying baselines in the second set of stimuli were computed linearly on the logarithmic semitone scale, because "the slope of the baseline is usually measured in terms of semitones per second" (Terken 1991, p.1772).

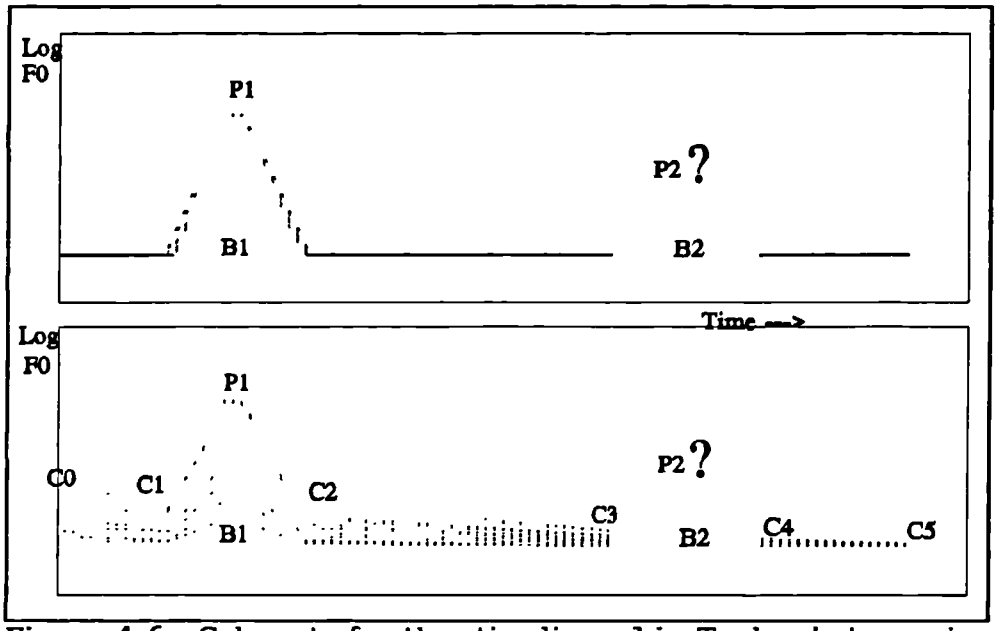


Figure 4.6 - Schemata for the stimuli used in Terken's two pairs of matching experiments. Top figure : stimuli for first pair (flat baseline). Bottom figure : stimuli for second pair (declining baselines).

If the hypothesis that relative prominence (or, respectively, relative pitch judgment) depends on difference in peak pitch (the MAX hypothesis) is correct, then the subjects should choose a value for P2 which was the same as that of P1, in all cases. If the hypothesis that relative prominence (or pitch judgment) depends on difference in pitch distance from peak to baseline (the CHANGE hypothesis) is correct, then the subjects should choose a value for P2 which was the same as that of P1 in the first pair of experiments (with the flat baseline), and a value for P2 for which the peak-baseline distance was the same as that for P1 in the second pair of experiments.

Terken found in the first pair of experiments (with the flat baseline) that values of P2 chosen to match P1, in the lower range values of P1, were equal to or a little higher than P1. This was also true of some of the mid-range values for the pitch-matching task. For the mid and upper-range values in the prominence-matching task, and the upper-range values in the pitch-matching task, values of P2 chosen to match P1 were significantly lower than P1, and more so for the prominence-matching task than for the pitch-masking task.

Table 4.1 - Values of Salient points for the eleven stimuli in Terken's second experiment-pair. (P1 values as given by Terken and same as for the first pair; the rest estimated by author from baseline spec. and Terken's (1991) Table I data).

C0	B1	P1	C1	C2	C3	B2	P2	C4
122	110	156	114	106	88	85	136	82
116	106	152	109	102	87	84	132	81
111	102	145	105	99	86	83	129	80
105	98	141	100	95	84	81	127	79
100	94	135	96	92	82	80	122	79
96	91	130	93	89	82	80	121	78
91	87	125	89	86	80	79	117	77
87	84	120	85	83	79	78	114	77
83	81	116	82	81	78	77	111	77
79	78	112	78	78	76	76	107	76
75	75	108	75	75	75	75	104	75

In the second pair of experiments (with the baselines of varying degrees of slope) the values of P2 chosen to match P1 were less than or equal to P1 only for the low-range values of P1 in the pitch-matching task. Otherwise, the P2 values chosen were always less than P1, increasingly so for higher-range P1 values, and significantly more so for the prominence-matching task than for the pitch-matching task. The results for the prominence-matching task are summarised in Figure 4.7 (redrawn from Figure 7 in Terken 1991).

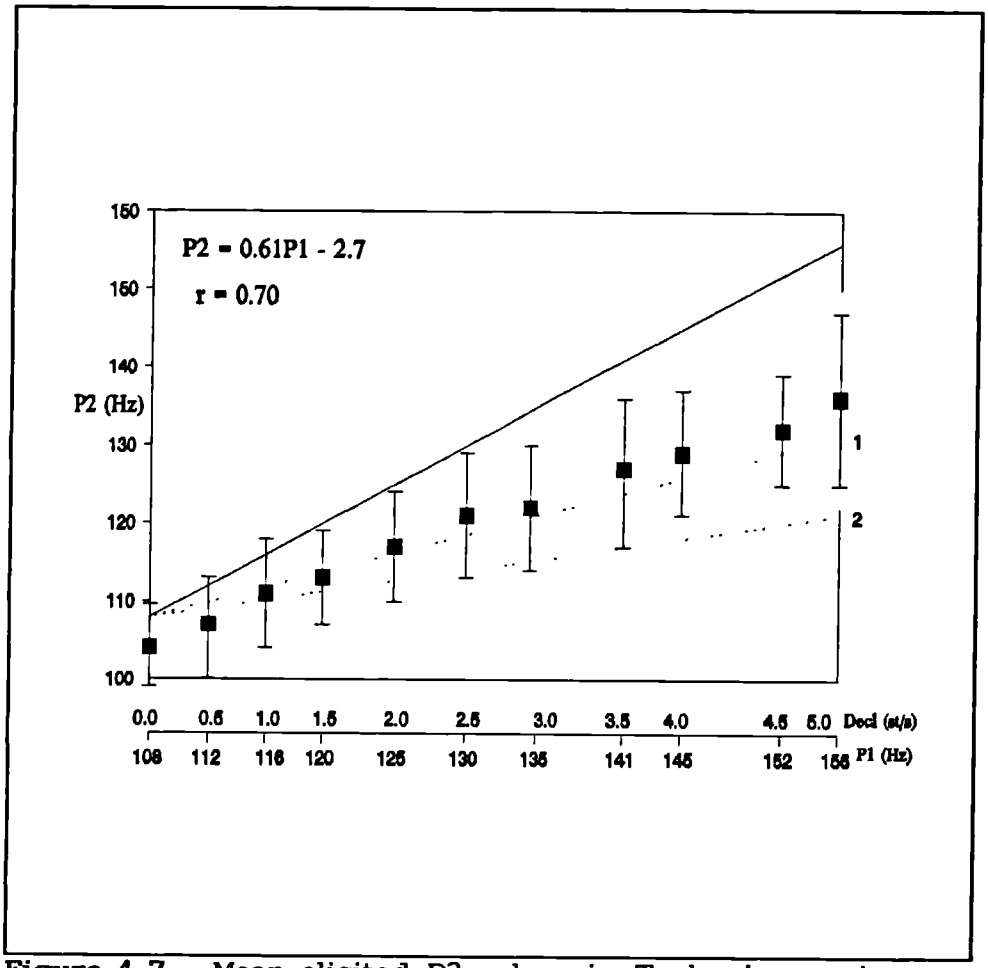


Figure 4.7 - Mean elicited P2 values in Terken's prominence-matching task with declining baselines. Solid line is expected behaviour in MAX hypothesis. Dashed Line 1 is that expected in CHANGE hypothesis on Hz scale, 2 is same on semitone scale.

Terken makes the following conclusions in the light of the results from the experiments. Firstly, he is able to reject the MAX hypothesis on the grounds that in all four conditions (Flat baseline+pitch-matching, Flat baseline+prominence-matching, Variable baseline+pitch-matching and Variable baseline+prominence-matching), the regression line of the elicited P2 judgments on P1 was significantly different from the P2=P1 line. Next, he feels able to say that the CHANGE hypothesis should be rejected. Although the regression line through the values of P2 which would support the hypothesis of prominence deriving from distance in Hz. from peak to baseline (dashed line 1 in Figure 4.7) is pretty close to the apparent trend through

the actually elicited P2 values[7], a regression line fit to those actually elicited values yielded a regression equation of
D2 = -14.99 + 1.39*D1, where D1 is the drop from P1 to B1 (see Figure 4.6) and D2 the drop from P2 to B2. The slope coefficient of 1.39 is fairly close to the value of 1 required for full support of the CHANGE hypothesis, but the values in the lower and upper thirds of the range militate against accepting it completely.

Terken also notes that the results from the first pair of experiments, when the baseline was flat, show evidence against the CHANGE hypothesis as well as the MAX hypothesis. The values elicited for P2 matching P1 in the middle and top of the range are less than P1 itself, even though the value of D1 is the same as the value of D2, and across the whole of the range.

This leads him to conclude that the declination effect is partly attributable to expected topline declination (we can call this the 'abstract part', or the 'expected declination that listeners compensate for', that Leroy (1984) referred to). This effect is operative both in the case of no declination in the baseline, and when there is such declination. However, the declination effect is much stronger in the latter case, which allows him to conclude that it is also partly attributable to actually occurring baseline declination (we can call this the 'concrete part', or the 'physically present declination').

He then goes on to argue that the baseline and topline should be taken to vary independently, and that topline variation makes the stronger contribution to variation in prominence. The argument is that the range of variation of the slope of the topline for equal prominence judgments is less than that of the slope of the baseline (this assertion presumably being based on the more finely varying results of the first pair of experiments, which are seen as evidence of what I have called the abstract part of the declination effect). Therefore, variation in the pitch values of the accent peaks (which are the carriers of the topline), independent of baseline variation, is going to have a correspondingly stronger influence on the relative prominence of

---

[7] The equivalent hypothesis in which peak to baseline distance is expressed in semitones being rejected outright, as witness the regression line through the P2 values which would support such a hypothesis (dashed line 2 in Figure 4.7).

the accents. At the same time, there will be some baseline slope variation which is not independent of topline variation, and that will also have a modulating effect on the relative prominence of the accents. As Terken puts it, "prominence relations should be accounted for primarily as a function of relations between F0 maxima, but the precise modelling of this function depends on the slope of the baseline" (Terken 1991, p.1775).

This view is taken to support a model of intonation consonant with that of Thorsen (see Chapter 2), in which an independently varying baseline component interacts with a local accent target component to produce an actual intonation contour. He also takes it as evidence against models of intonation based purely on sequences of accent targets ('tone sequence' theories of intonation, as in Liberman and Pierrehumbert 1984), since the baseline has been shown to be an essential part of the model of intonation for the purposes of relative prominence modulation, and the baseline is "a global property of the intonation phrase" (Terken 1991, p.1775).

4.2.3.1 Discussion

Terken's work is interesting, because it aims to discover how the different parts of the declination effect that Leroy referred to can be attributed to different components of the intonation contour; it is a step on the path to a more detailed 'anatomy of declination'. Some of the effect is attributed to compensation for expected declination in the topline, and some to physically present declination in the baseline.

There is one problem with Terken's partition of the declination effect, which can be just mentioned before continuing. He appears to conclude, from the difference in the strength of the declination effect between the first and second pairs of experiments, that the effect of compensation for expected topline declination seen in the first pair is preserved in the second pair and added to an effect attributable to baseline declination. He is not completely justified in concluding this; it may have been that the effect he identifies as attributable to expected topline declination only occurs in the absence of baseline declination[8], and that when there is baseline declination, all the

[8] This possibility seems more likely if it is remembered that the flat baseline is at quite a low pitch, not typical for the double pointed-hat pattern shown. This point will be addressed further below.

declination effect that is accounted for in the analysis is attributable to the baseline declination.

He is perhaps also not justified in concluding that his results corroborate those of Leroy, that "[f]or flat baselines, the expected topline is steeper than the baseline, and for steep baselines the topline is less steep than the baseline" (Terken 1991, p.1773), for the simple reason that the marginal significance of Leroy's results in this respect can be seen to be even more attenuated after reanalysis (see footnote 1).

Terken maintains in his concluding remarks the notion of an abstract global construct, the topline, through which compensation for expected declination is mediated as part of the declination effect, along with another global construct, the baseline, through which 'physically present' declination is mediated as part of the declination effect. He includes a set of local constructs, the pitch accents, and presumably proposes that the prominence of these is framed against the global topline and baseline. So he departs from the standard Eindhoven approach in which accent peaks are fixed to a topline which declines more steeply in the linear frequency domain than the baseline to which their bases are fixed, and instead proposes a topline which declines less steeply than the baseline, but which is more abstract than the baseline in that accent peaks of differing prominence don't need to occur on it.

Yet the division of labour in framing prominence between physically present declination on a more concrete baseline and 'expected' declination on a more abstract topline is not without problems. In the first place, there are many contours in which the topline is more concrete than the baseline, when modelled using straight-line stylisations (such as a single 'flat-hat' contour - see Chapter 2). Secondly, and more importantly, the stick which Terken uses to beat the 'Tone Sequence' theories of intonation - the globality of the baseline needed to frame the prominence of accents - can be turned against him; for the contribution of the baseline to the framing of accent prominences is in the declination that is physically present in it.

It is the mark of the globality of an intonational entity that it have some degree of abstraction. The baseline in Terken's experiments can be considered to be partly abstract, in the sense that it doesn't coincide with

the actual F0 contour during the accented syllables[9]. Yet the purpose it serves, according to Terken, is to modulate the prominence relations between the accented syllables based on their peak pitch with the degree of declination which is physically present, or which is, to use the term Terken uses, "audible" (Terken 1991, p.1773). If that is so, why could the same effect not be mediated by the declination audible in local stretches of the intonation contour? To put it another way: Terken has suggested that the declination effect is the result of a more 'abstract' declination component combined with a more 'concrete' component, but that both components are global properties of an intonation contour. Could it not, however, be the case that the concrete component of declination is the result of purely local intonational processes? There is nothing in the results we have seen so far which suggests this might not be the case.

4.3 DECLINATION HYPOTHESES - LOCAL VS. GLOBAL

In order to investigate this question further, it is useful to set up this hypothesis incorporating only local declination against one incorporating global declination. Thus far, there is justification for doing so. Terken's abstract component of the declination effect, motivated by the results of his flat baseline experiments, can be called into question if it is considered that the baseline was chosen at a level of 75Hz, low enough in virtually any speaker's range for a major tone-unit boundary to have been perceived as existing between the first and second accents. In this case, the declination effect that he observed could have been the result of a phrasal downscaling of pitch range at a tone-unit boundary, an even stronger possibility when it is remembered that reiterant speech was used in the experiment, so that no textual cues to phrasing would be available. Moreover, the PSE in Leroy's condition 'M' (with the flat baseline), predicts only a small 'abstract' declination effect, and in any case doesn't look to be significantly different from 0 (though she didn't test for this). Both these facts tie in with the fact that she used a higher flat baseline (of 90Hz) than Terken. Both facts are also consonant with the hypothesis that any declination effect is attributable only to some combined function of the individual local stretches of declining intonation. Simlarly, whatever the true interpretation of Gussenhoven and

---

[9] Here it is being assumed that a treatment of the F0 contour as a single-component signal is valid, and operative.

Rietveld's results, there is no evidence that the declination effect observed could not have been a function of local declination.

By the same token, it is possible that the results in all three sets of experiments are attributable to a declination effect due to an internalised abstract declination function, which is operative in all utterances, but variations of which can be cued in an as yet unspecified way by 'physically present' declination. This is a more general statement of the position that Terken reaches that physically present declination modulates the effects of an abstract declination function; the latter may consist of an abstract declining topline, or baseline, or both. Again, Leroy's position was that physically present declination only partly contributes to the declination effect, the rest being attributable to a necessary abstract global declination function. And Gussenhoven and Rietveld's results left open the possibility that the declination effect they elicited was partly due to the physically present declination and partly due to an abstract declination function (which they unwarrantedly attributed to Final Lowering).

It is thus useful to articulate two competing hypotheses, a Global Declination Hypothesis and a Local Declination Hypothesis. These hypotheses are taken to be valid whether the declination effect is attributed wholly to declination per se or whether it is partitioned into the effects of Initial Raising, Final Lowering and the optional downstep, as discussed in Chapter 2, along with a reduced declination function proper.

The two hypotheses both make the following assumption : that declination has the effect of increasing the prominence-F0 ratio on consecutive accents in a major tone-unit (along, perhaps, with Initial Raising, Final Lowering and Downstep). This assumption has the following corollary : Consecutive accents in a major tone-unit have equal prominence if they exhibit a decrease in their peak pitch, conforming to a particular (possibly piece-wise) function.

GDH (Global Declination Hypothesis)

PART 1

Declination consists of a global function which does not necessarily correspond to physically manifested declining F0 in an intonation contour.

It can be modelled as a declining baseline (of >= 1 piece), as a declining topline (of >= 1 piece) or as a declining topline and baseline (of possibly different slope) above, below or within which intonation contours are scaled).

PART 2

There is a global declination effect such that consecutive accents in a tone-unit have a prominence-F0 ratio which increases with respect to that of the first accent, regardless of the intervening material.

COROLLARY

Consecutive accents in a tone-unit have equal prominence ONLY if they exhibit a decrease in their peak pitch values, conforming to a particular function, ceteris paribus.

LDH (Local Declination Hypothesis)

PART 1

Declination has to be physically manifested in some part of the intonation contour to have an effect on the prominence ratings elicited for following accent peaks.

(This hypothesis can be interpreted as implying that a nominally level stretch of pitch which is, in fact, gradually declining must occur between the inner bases of consecutive accent peaks for there to be a declination effect on the second peak. The nature of this declination effect is described in the second sub-hypothesis.)

PART 2

There is a local perceptual declination effect, such that if a local declination is present between two accents, it increases the prominence-F0 ratio on the second accent.

COROLLARY

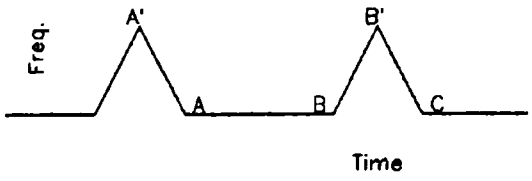
If there is no local declination between two equally high accent peaks, the two accent peaks are of equal prominence, ceteris paribus.

In the corollaries to these two hypotheses, the rider 'ceteris paribus' expresses the point that the relative prominence of accents within the accented area could also be affected by the shape of the contour outside the accented area. The corollaries are then only valid if that factor doesn't act to make the prominence of the respective accents unequal, notwithstanding the shape of the contour in the interaccentual stretch. The question of the effect of the shape of the contour outside the accented area is taken up in Chapter 5.

LDH requires there to be no prominence difference between accents of the same peak F0 separated by a level stretch of F0 (not so low that major tone-unit boundaries are perceived) of zero slope. As there is no local declination in this case (and the prominence-F0 ratio is changed otherwise only by downstep, initial raising or final lowering, which aren't operative here), there can't be any change in the prominence-F0 ratio, so that equal F0 gives equal prominence.

4.4 An experiment to test GDH against LDH

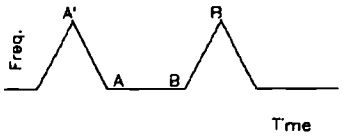
The most direct way of testing LDH against GDH would be to determine the prominence ratio, as elicited from listening subjects, between the equally high accent peaks A' and B' in the following contour:



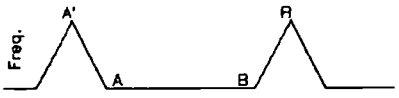
in which the value of F0 at point A is the same as the value of F0 at point B, and is not too close to the bottom of the speaker's range. Under the GDH, B' should be rated more prominent than A'; under the LDH, they should be rated the same. This could be done by presenting subjects with a set of stimuli with a range of values on B' varying around a central value equal to that on A' and asking them in each case either whether B' was less prominent than A' or more prominent than A' or what prominence value (on an

appropriate scale) they would assign to A' and what to B'. However, it was felt that this task was quite similar to that posed in Leroy's condition 'M', and that while a replication of that experiment would have some worth, it would not have a chance of revealing any unknown hidden effect bearing on the issue (for instance, variation in the distance between A and B). For the same reason, a prominence-matching task similar to Terken's was not performed (although there would have been more gain in doing that, with a baseline value less conducive to the evocation of an interaccentual tone-unit boundary).

The first thing that was done to gain some new insight into the respective merits of the two hypotheses was to test prominence judgments on a token with one interaccentual distance against another with a longer such distance. That is,



was tested against



Assuming that the extra distance between A and B in the second stimulus was perceptually detectable, this paradigm necessarily gave the capability of comparing LDH against GDH, because under LDH the pitch heights[10] of accents A' and B' are identical in both cases (they are computed relative to the physically present baseline), whereas under GDH, (i) the pitch heights of A' and B' are different in both stimulus types (because they are computed relative to a declining reference line) and (ii) the difference in pitch height between A' and B' is most probably greater in the second stimulus type than in the first. That is, as the duration between A' and B' increases, so, it is most likely, does the value (pitch height of A') - (pitch height of B'). This is true whether the declination function is

[10] 'pitch height' will be used to refer to the perceived pitch of a point in the physically manifested contour relative to a declining baseline, be it concrete or abstract.

(A) linearly time-independent (and so has variable slope and constant drop with time) or is

(B) linearly time-dependent (and so has constant slope and variable drop with time) or is

(C) non-linearly time-dependent (and so has variable linear slope and variable drop with time).

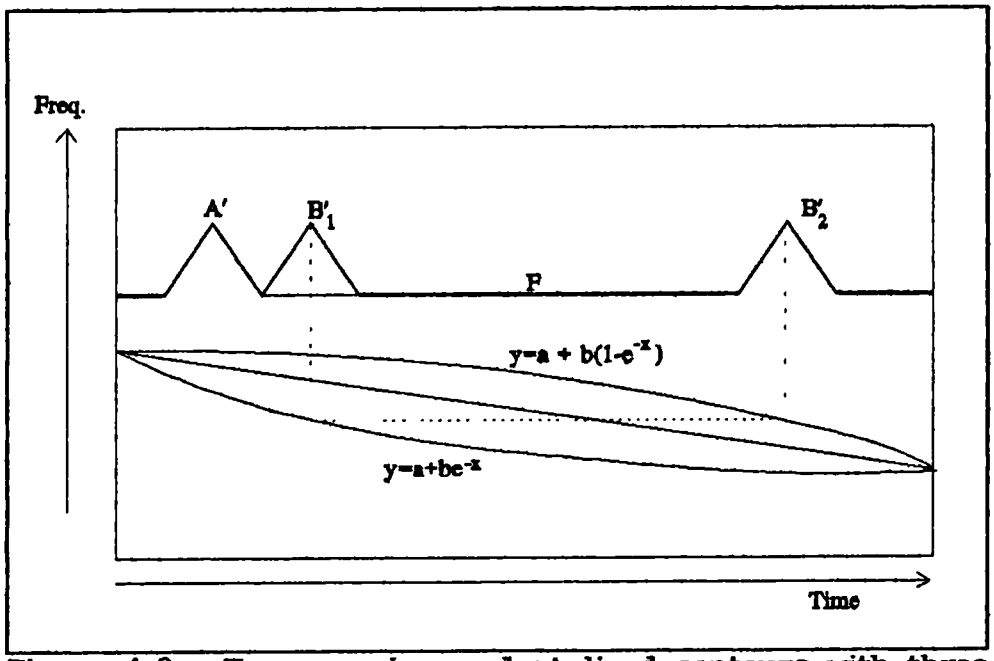


Figure 4.8 - Two superimposed stylised contours with three different declining reference lines. Contour 1 = A' + B'$_1$ + flat baseline to point F. Contour 2 = A' + B'$_2$ + all of flat baseline.

Cases (B) and (C) can be seen in Figure 4.8.

Case (A) holds because the proportion of the total duration of the utterance between the beginning and the first accent and between the second accent and the end decreases as the distance between A and B increases, and the 'amount' of declination between A' and B' thereby increases.

It is stated above that it is only probable that the difference in pitch height between A' and B' is different in longer utterances than in shorter utterances. This is because there are two cases in which the value (pitch height of A') - (pitch height of B') could remain constant as the duration between A' and B' varies :

(i) The declination function is non-linearly time-dependent, but is a different function in the case of the short test schema above than in the case of the longer test schema above. For instance, when the distance between accents is small, the declination function might be of the form

1. $a + be^{-x}$

(the bottom reference line in Figure 4.8), whereas when it is large, it might be of the form

2. $a + b(1-e^{-x})$

(the top reference line in Figure 4.8). In this case, the pitch height of the two putative second accents might be identical (as marked by the dotted lines in Figure 4.8), so the value (pitch height of A')-(pitch heght of B') could remain constant in these circumstances. However, it is clear there would need to be some special, independent evidence for such a difference in the form of the declination function between cases where accents are widely separated and those where they are narrowly separated. One possibility would be that cricothyroid muscle relaxation (as represented by declination function 1 above) is used as the physiological analogue of the declination function in short utterances, whereas subglottal pressure decline (as represented by the second declination function above) is used as its physiological analogue in long utterances[11]. The speaker might then choose one or the other physiological process as the base function from which the baseline is computed, depending on how long he or she expected their utterance to be, and the hearer would choose the appropriate function for scaling the prominence of the accents on the strength of the length of the utterance once they'd heard it.

(ii) the declination function is partly time-dependent and partly time-independent. That is, as the distance between A and B increases, so the frequency drop between the start and end of the declination line increases, but not so as to preserve the same slope (the absolute value of the slope decreases). In this situation, the value (pitch height of A')-(pitch height of B') might happen to be the same for different values of the distance between A and B.

---

[11] Collier (1987) proposes such a dual physiological analogue for declination, but within the same utterance (cricothyroid muscle activity being the main controlled variable at the start of the utterance and subglottal pressure decline being the main one at its end).

In both these cases, the form of the declination function is taken to be chosen at the start of the utterance. Therefore, in order to control for the possibility of such an effect, an additional modification was added to the stimulus set, viz. the interruption of the contour immediately after the second accent and the addition of a filled hesitation pause. This was intended to provide the listener with a cue for a declination function which was appropriate for a longer utterance than the one in which no modification took place. Thus, if the declination function used by speakers and hearers were of the form suggested in either cases (i) or (ii) above, the value (pitch height of A')-(pitch height of B') would be more negative in the case of the overtly 'finished' contour than in the case of of the overtly 'unfinished' contour. That is, the slope of the abstract declination line (mean instantaneous slope at the peaks of the two accents in the non-linear cases) would always be steeper in the case of the finished contour than in the case of the unfinished contour.

If the two conditions are combined, there are four stimulus types: short finished, short unfinished, long finished and long unfinished. For equal pitch on the peaks A and B, one would expect higher prominence on B than on A in all these cases if the GDH were true. At the same time, there should be more difference in prominence between the two accents in the case of the long contours than in the case of the short contours, except in the case when the declination function is not simply time-dependent or time-independent but is proportionately so depending on the expected duration of the utterance, in which case the prominence between the two accents could be the same in each case of the long contours as in the corresponding cases of the short contours. In this case, there should be more difference in prominence between the two accents in the case of the finished contours than in the case of the unfinished contours. On the other hand, if the LDH were true, the difference in prominence between B and A should be the same for all stimuli, and should be zero.

Looking at Figure 4.8, it is possible to protest that the proposed stimulus set set is biased against the GDH because there are no cues at all to declination, thus making the stimuli unnatural, and unlikely to cue a baseline declination function in the way envisaged. At this point, it could be said that in the form stated, the GDH does not require there to be any cues for the abstract

declination line (which, it should be emphasized, could equally well be a topline as a baseline). Futhermore, there might in any case be cues to the declination line in variation in voice quality, as suggested by Pierrehumbert (1980, p.136).

However, in order to make the contours more natural in this respect, the decision was taken to include an additional modification to provide an additional cue to a putative baseline, viz. a ramp down in the final section of contour (after the second accent, so that the final F0 value was lower than the initial one). This constituted a 'physically present' final lowering over the last half-second of the contour. Now, in the case of utterances of the same length, it would be impossible to tell whether any difference this made in the rating of the prominence on B was due to a local final lowering effect or a global declination line cued by the start and end F0 values. In the case of utterances of different length, though, any difference in relative prominence on B due to the cued declination line would show up separately from any such difference due to final lowering. That is, the effect on elicited prominence on B due to local final lowering would be the same for long and short utterances, but the effect due to a global declination function should be greater in longer utterances than shorter utterances[12].

There were thus three variations on the basic intonation pattern of two accented syllables occurring between level stretches of pitch. The first variation was prolongation of the level stretch between the two accented syllables. The second variation was curtailment of the level stretch after the second accented syllable and insertion of a filled hesitation pause. The third variation was the introduction of final lowering during the last half-second of the final level stretch. The second and third operations required mutually exclusive operations on the basic intonation pattern, but the first was independent of the other two. This meant that the first variation could be applied to the basic pattern and to the second or third variations of the basic pattern, yielding six patterns in all, as seen in Figure 4.9.

---

[12] That is, unless either of the cases mentioned above holds, in which the value (pitch height of B)-(pitch height of A) is the same for particular duration values of long and short utterances. These cases could not be distinguished by the experiment, because it is in principle impossible to make the final-lowering modification to the unfinished stimuli (with filled hesitation pause).

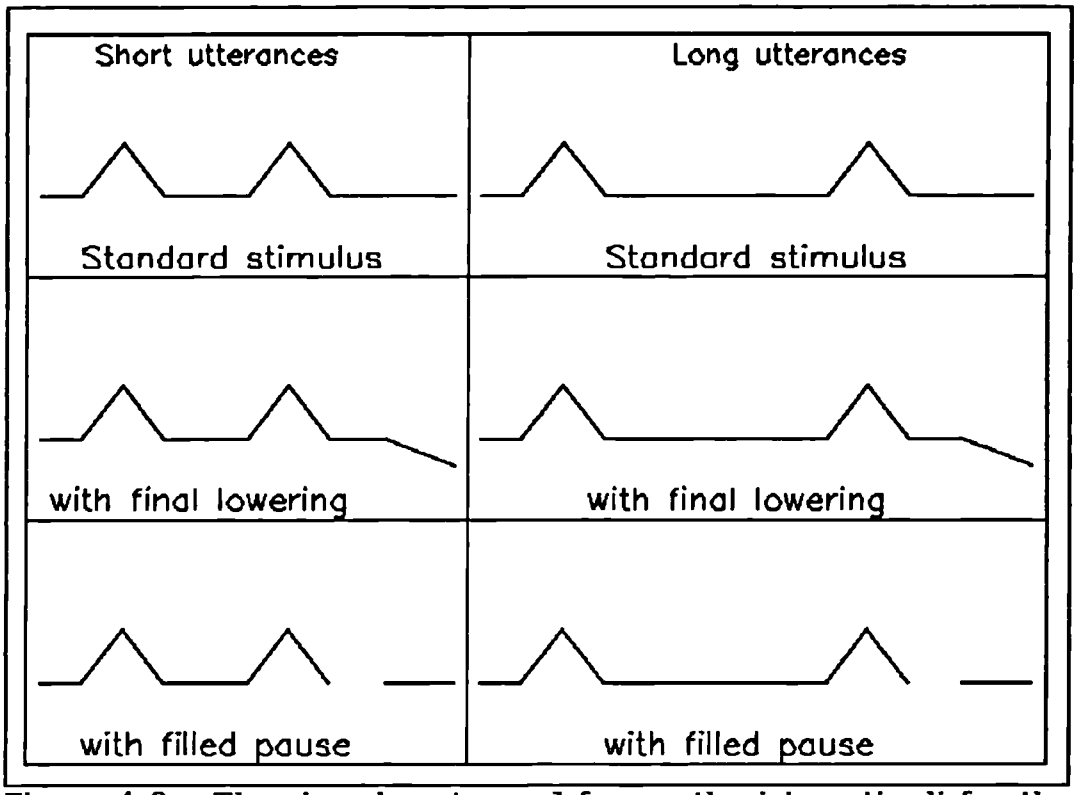


Figure 4.9 - The six schemata used for synthesizing stimuli for the perceptual test

4.4 AN EXPERIMENT TO TEST THE GDH AGAINST LDH

4.4.1 Stimulus Preparation

The experiment was designed, prepared and executed whilst on ERASMUS scheme study-leave in the Phonetics Institute at the University of Nijmegen in the Netherlands. Stimuli were prepared by resynthesis using a PSOLA[13] algorithm developed and encoded by Robert Espesser at the University of Aix-en-Provence.

In preparing the stimuli, provision had to be made for the slope of the abstract declining reference line being quite small (since some of the declination effect could have been taken up by Final Lowering). Therefore, it was considered that a fair degree of accuracy was required in the preparation of stimuli, and that effects such as variation in F0 resulting from segmental coarticulation should be carefully controlled for. In this latter

[13] Pitch-Synchronous OverLap-Add analysis and resynthesis of speech; see Hamon, Moulines and Charpentier 1989.

regard, the minimal degree of control considered necessary was the use of identical segmental material in the accented syllables.

A normal sentence was chosen for the basic stimulus, rather than reiterant speech, so that textual phrasal cues could be used to discourage the interpretation that a major tone-unit boundary could be placed between the accents. The basic sentence to be used in the experiment was spoken in Dutch, the perceptual experiment being performed with Dutch subjects as listeners. The following sentence pair was devised which satisfied the requirement of segmental consistency :

Z1 : Van paJAREn maken we na JAREn limonade
We produce lemonade from pajaars, after a number of years(' preparation).

Z2: Van paJAREn maken we in Leeuwarden na JAREn limonade
In Leeuwarden we produce lemonade from pajaars, after a number of years(' preparation).

(Note that 'Z' stands for 'zin', which means sentence. In these sentences, capitalised syllables are accented. Note that "limonade" is deaccented). The word 'pajaren' was constructed specially for the experiment, being a plural form referring to an imaginary tropical fruit, the 'pajaar'. Obviously, recourse to such innovation was motivated by the fact that the accented syllables in pajaren and jaren were to be identical in segmental structure.

In addition to controlling for segmental structure, a decision had to be made about the frequency quantity used. In this regard, the study by Hermes and van Gestel (1991) was relevant. In it, an experiment was conducted (in which cross-register pitch-matching tasks were performed on three basic intonation contours: rise, rise-fall and fall, uttered using reiterant speech) which led to the conclusion that an ERB-rate scale most closely matched the scale used by listeners in intonation-processing tasks, and that ERB-rate should be considered the quantity varied by speaker-hearers when communicating by intonation.

Hermes and van Gestel's experiment does appear to substantiate that claim, but since there is the possibility of subjects listening in different ways (e.g. musically vs. non-musically), and since the non-declining contours of the proposed experiment might be seen to favour a musical mode of listening (for which, as Hermes and van Gestel claim, the semitone scale is favoured), their experiment should better normally be repeated using the stimuli to be used in the main experiment, in order to check on the frequency scale used.

In the event, this was not done, because the current experiment did not involve variation of F0 (except in the stretch undergoing final lowering). However, the problem of determining which mode of listening a listener is using certainly needs to be addressed, and one way of doing this is to precede the preparation of stimuli for a particular experiment with a Hermes and van Gestel -type test in which the perception of the object stimuli resynthesized using various frequency scales is compared.

It was felt that the other major prosodic variables, duration and amplitude, should also be controlled for during the accented syllables. Prior to performing the experiment, it was considered that the easiest way of maintaining the same amplitude, duration and F0 in the accented syllables was to edit in the same syllable [ja:] in both positions in sentences Z1 and Z2. Thus, the stimuli for the experiment were prepared as follows:

(1) The sentences Z1 and Z2, spoken by a male speaker (TR) with an intonation approximating that of the basic stimulus, were recorded onto reel-to-reel audio tape in a professional-style recording studio. These were acquired onto a VAX mini-computer at a 10kHz sampling rate.

(2) One instance of Z1 was chosen as the basic token (Stimulus 1), from which the five other stimuli were created. The dual-accented contour to be used for stimulus 1 was generated, and stimulus 1 synthesized at the same time. The level (physically present) baseline of the contour was set to 112Hz (= 3.75 ERB, using Greenwood's formula as published by Hermes and Van Gestel 1991), and the peak of each accented syllable to 142Hz (= 4.5 ERB). The rise time of each accent was 100 ms, the fall time 100 ms, and the duration of the flattened peak of the accent 50ms.

(3) The first accented syllable was pasted in place of the second, in the following way:
for both [ja:] syllables in the original resynthesized token, the left boundary was determined at a zero-crossing in the middle of the [j] of the syllable [ja:], the right boundary similarly in the middle of the following [r]. The second accented syllable was then excised; the first was cut out, replaced in its original position and pasted in place of the excised second accented syllable. Also, the amplitude of the signal immediately to the right of the second accent was reduced to match more that appearing in the pasted syllable's natural right context. Replay of the resultant contour revealed there to be no audible discontinuities in the resultant token, which sounded quite natural. This token constituted the first stimulus. Figure 4.11a shows the speech pressure waveform and F0 trace for stimulus 1, Figure 4.11b the position in the waveform of the starts and ends of the accent in the [ja] syllable.

(4) The second stimulus, which was the same as the first but with final lowering, was generated using a fundamental frequency editing program. The end point of the contour was reduced from 112 Hz to 99 Hz (3.4 ERB), and linear interpolation performed between that point and a point 500 ms prior to it. PSOLA-resynthesis was then performed. This produced a natural final-lowering effect.

(5) The third stimulus, which was the same as the first but with most of the last word excised and replaced by a filled hesitation pause, was created in the following way: The first stimulus was curtailed half-way through the [l]

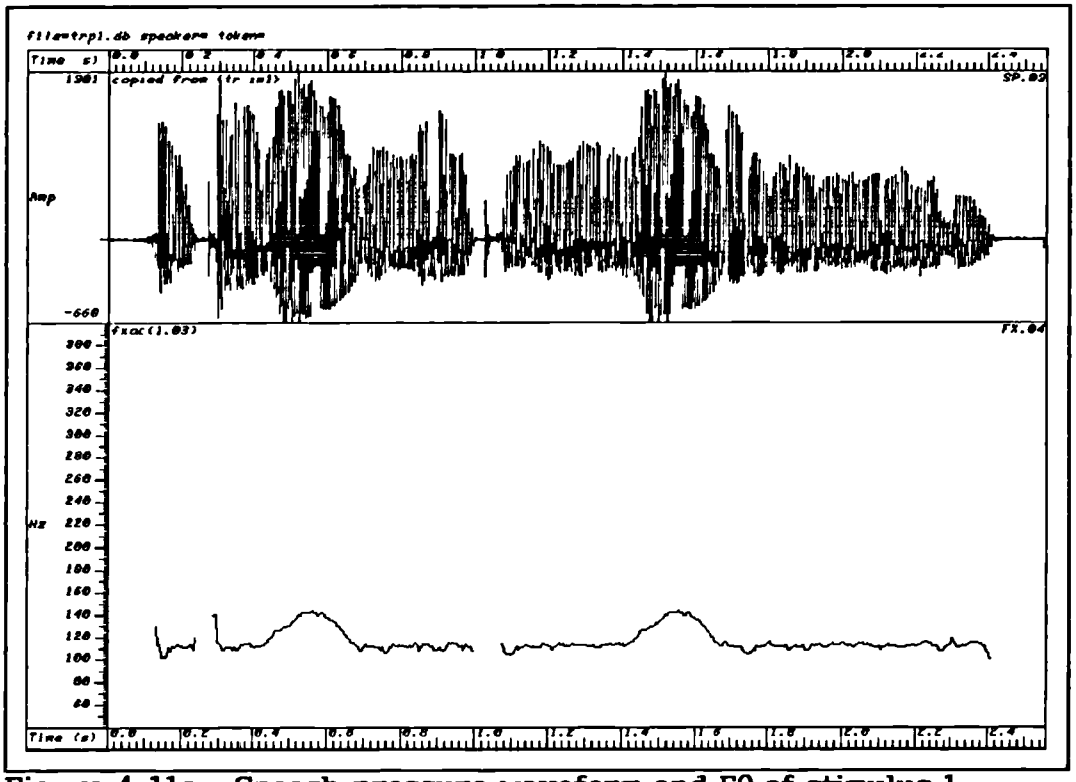


Figure 4.11a - Speech pressure waveform and F0 of stimulus 1

of "limonade", and the amplitude of the trailing bit of [l] ramped down to a tenth of its previous amplitude. A silence was then inserted of 500 ms length, and a filled hesitation pause (of Dutch type, which is a rounded central vowel) of 1.24 s length pasted from a separate file. This filled hesitation pause had been previously been recorded at the same time as the original sentences Z1 and Z2, and resynthesized with a level value of F0 of 112Hz, the same as the baseline in the basic token. Note that the inserted silence simply consisted of a stretch of zeroes, which was perceptibly different from the ambient (very low-level) noise perceivable elsewhere in the tokens when listened to over headphones. This problem of unnaturalness was overcome in the experimental situation by playing the tokens to subjects through a loudspeaker, under which conditions the difference in the silent stretches was not apparent, because of the masking effect of the ambient noise in the listening studio.

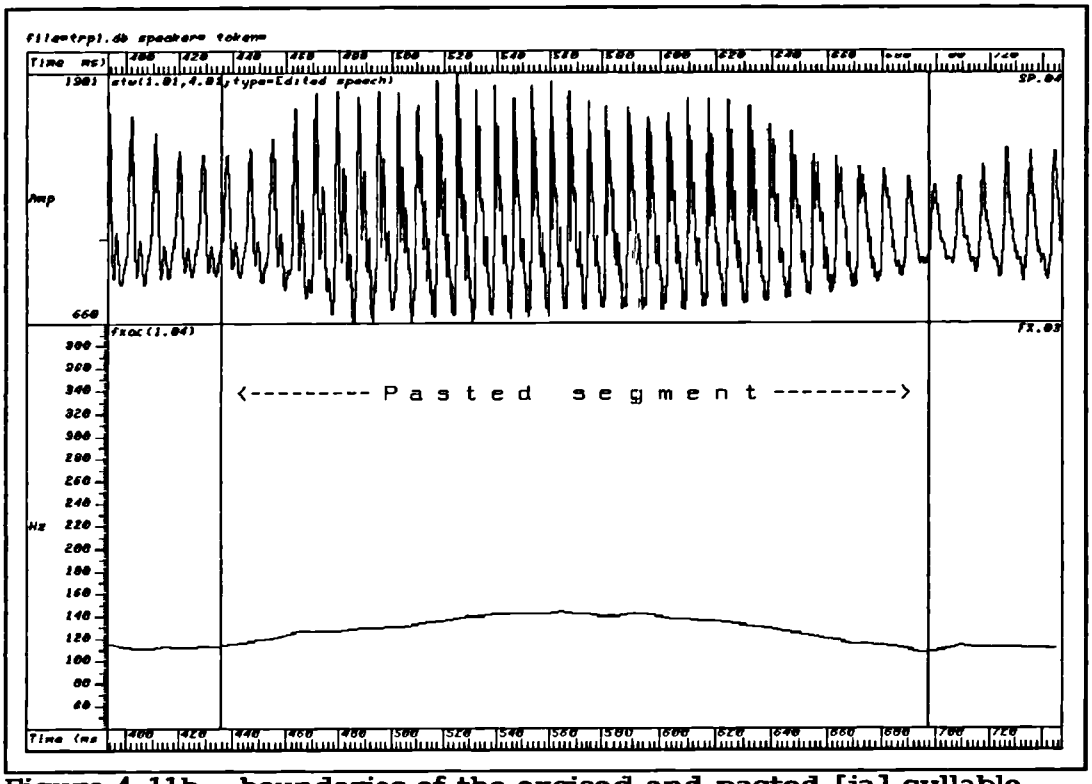


Figure 4.11b – boundaries of the excised and pasted [ja] syllable

(6) The fourth to sixth stimuli were the same as the first to third, but with the phrase "in Leeuwarden" pasted from Z2. These (unaccented) words had also been resynthesized with a level contour of 112Hz. The cutting and pasting was again performed at zero-crossings.

The preparation of the six stimuli is summarized in Figure 4.12[14].

4.4.2 The perceptual experiment

The GDH was to be tested against the LDH primarily by comparing the shorter with the longer stimuli. If the relative prominence assigned to the second accented syllables compared with the first accented syllables was greater in the longer stimuli than in the shorter, then there was corroboration for the GDH. One possible task for the subjects to be presented with to determine these results was for them to rate on an appropriate scale the prominence of the first accents (which would remain fixed) and the prominence of the second accents (F0 values of which in

[14] "STW", "STIMGEN" and "FXEDIG" are all names of programs adapted or developed by the author.

different stimuli would vary around a central value equal to the value on P1). Comparisons of the PSEs on the labelling curves determined from these

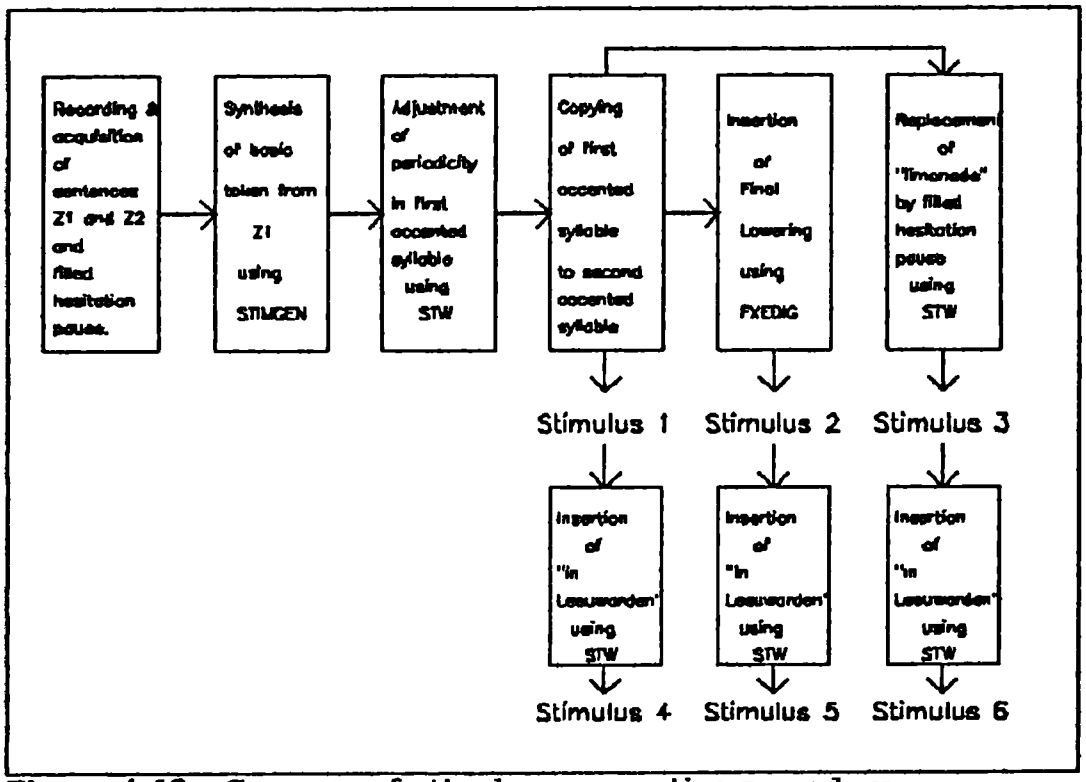


Figure 4.12 - Summary of stimulus preparation procedure

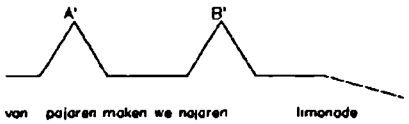
results between the six different stimulus types, would then decide between the LDH and GDH.

However, it was felt that the number of stimulus presentations resulting from such a task (say 6 (for each of the stimulus types) x 11 (for each of the F0 values on B tested) x 3 (for repetitions of individual stimuli) = 330, for each of which two prominence judgments would be required) would be too fatiguing for subjects.

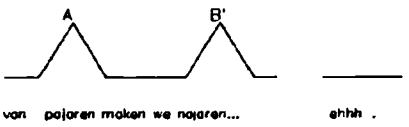
At the same time, a prominence-matching task, as in Terken's experiments, was rejected, because it was desired to give subjects not too much time to make a decision, to prevent over-analytical judgments taking place, perhaps approaching comparative judgments of the pitch of the accented syllables.

Therefore, the basic task that subjects were presented with was to compare the relative prominence of the second accent in one utterance with that of the second accent in another utterance[15], on a comparative seven-point scale. The fact that this was a paired comparison task meant that the number of stimulus presentations could be considerably reduced without affecting the significance of the results, which meant that the task would not be too fatiguing for the subjects.

Thus, comparative prominence ratings were elicited for the second accented syllables in pairs of stimuli. For example, for the stimulus pair



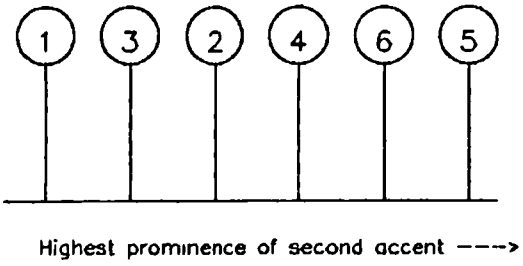
vs.



subjects had to compare the prominences of the B' accents, stating whether they were considered equal, or whether the second was more prominent than the first (on a 3-point scale) or whether the first was more prominent than the second (on a 3-point scale). These comparative ratings were then to be subjected to a type of ANOVA developed specifically for the analysis of difference scores (Scheffe, 1952). This method yields a set of scale values for each object (in this case stimulus type), and a 'yardstick' value which enables the experimenter to determine whether the separations between objects on the computed scale are significant. The kind of scale values that might be expected if the GDH were true would be as follows:

---

[15] They weren't asked to compare the difference in prominence between the first and second accents of one utterance with the that of another, as it was felt that this would have been too difficult a task for subjects.



Thus, the increased inter-accentual duration in stimuli 4-6 should separate them from stimuli 1-3 on the scale, because the partly time-dependent (or time-independent) putative reference declination line should be at a lower frequency value at the point that the second accent is reached in the longer stimuli (assuming the special conditions mentioned above don't hold), so the second accent should be perceived as more prominent. At the same time, the final lowering/increased cue to a declining baseline present in stimuli 2 and 5 should, if effective, separate them from the other stimuli. Note that in the suggested result, the said effect of final lowering/increased baseline cue is not so great as that of the increased interaccentual duration, as stimulus 4 is scaled ahead of stimulus 2.

Finally, in case the global reference declination function is partly time-dependent and partly time-independent, or is early rapid-decaying in short utterances and late rapid-decaying in long utterances, the second accents in the finished utterances should be perceived as more prominent than the second accents in the unfinished utterances. So stimuli 1 and 4 respectively would be scaled ahead of stimuli 3 and 6. If this were not the case, they should be scaled in the same position.

4.4.3 Experimental Procedure

All possible pair-wise permutations of the six stimuli (6 x 5 = 30 stimulus pairs) were randomly recorded in three blocks of ten on reel-to-reel audio tape. Each stimulus was separated from its partner by a one-second pause. Pairs were separated from each other by a four-second pause, blocks by a

ten-second pause. Prior to the thirty stimulus pairs, five randomly selected such pairs were recorded, in order to familiarise subjects with the material.

Fourteen members of the Phonetics Institute at Nijmegen University were used as subjects. They were given instructions in Dutch as to how to perform the experiment (these appear in Section 4.6 Appendix 2 with an English translation), and performed it in a sound-treated booth, without headphones. It lasted no longer than fifteen minutes per person.

4.4.4 Results

Table 4.2 lists the pooled responses of the fourteen subjects for each sentence

Judgment Z1>>>Z2	Z1>>Z2	Z1>Z2	Z1=Z2	Z2>Z1	Z2>>Z1	Z2>>>Z1
Zinspaar						
(1 2)			6	7	1	
(2 1)		3	8	3		
(1 3)		1	10	2	1	
(3 1)		1	10	3		
(1 4)		1	8	4	1	
(4 1)	1	2	5	4	1	1
(1 5)		2	5	6	1	
(5 1)		2	7	3	2	
(1 6)		2	8	3	1	
(6 1)		5	4	5		
(2 3)			7	6	1	
(3 2)		4	9	1		
(2 4)		2	8	4		
(4 2)		3	7	4		
(2 5)			8	6		
(5 2)		1	7	5	1	
(2 6)			8	4	2	
(6 2)	1	3	5	5		
(3 4)		4	6	3	1	
(4 3)		4	3	5	2	
(3 5)		4	5	3	2	
(5 3)		1	5	6	2	
(3 6)		1	8	5		
(6 3)		3	7	3	1	
(4 5)		2	6	3	3	
(5 4)		2	6	5	1	
(4 6)		1	9	3	1	
(6 4)			6	7	1	
(5 6)	1	1	8	4		
(6 5)		2	7	4		1
Totals	3	57	206	126	26	2

Table 4.2 - profile of judgments made of each stimulus pair

pair. Table 4.3 is one possible collapsing of Table 4.2, in which all three conditions where Z1 is rated more prominent than Z2 are collapsed into one condition (Z1 > Z2), likewise for those where Z2 is rated more prominent than Z1 (Z2 > Z1), and the effects of ordering within pairs are disregarded. Thus, it was true for 20% of instances of stimulus 1 that the second accent was considered more prominent than the second accent of the stimulus it was paired with, but similarly true for 31% of instances of stimulus 3.

Application of Scheffe's ANOVA for difference scores yielded the following results:

1. There was no significant main effect. That is, in general, the prominence assigned to the second accent in a particular stimulus type was not significantly more or less than that of its partner ($p<0.01$, $F(5,390)=1.3$).

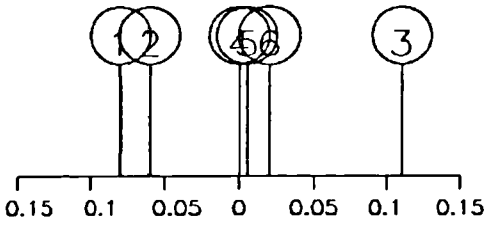
2. There was a significant effect of the ordering of the stimuli ($p<0.01$, $F(15,390)=4.04$). That is, in general, the fact that a stimulus appeared second or first in a pair had a significant effect on the judgment that its second accent was more prominent than that of its partner. (This can

Judgment Zin	Z1>Z2	%Z1>Z2	Z1=Z2	%Z1=Z2	Z2>Z1	%Z2>Z1
1	28	20%	71	51%	41	29%
2	29	21%	73	52%	38	27%
3	43	31%	70	50%	27	19%
4	40	29%	64	46%	36	25%
5	37	26%	64	46%	39	28%
6	37	26%	70	50%	33	24%

Table 4.3 - Table 4.2 reduced to three categories and irrespective of order of stimulus preparation.

clearly be seen from the totals at the foot of Table 4.2; there is a fairly strong bias of responses to the right of the scoresheet, indicating that if the second accent occurred in the second stimulus of the pair, it was more likely to be rated more prominent than that in its partner stimulus than if it occurred in the first of the pair).

3. The hypothesized scale underlying the judgments can be expressed as follows:



The distance between the lowest-rated object (stimulus 1) and the highest-rated object (stimulus 3) (= 0.20238) is not significant ($p < 0.05$) on this scale, using the yardstick value returned by the ANOVA for difference scores program ($Y_{0.05} = 4.03 * 0.0633 = 0.2563$). Since Scheffe states "The suggested test of the main effects is to declare them significant at the E level if and only if the largest and smallest of the estimated main effects a_i differ by more than the "yardstick" Y_E " (Scheffe 1952, p.392), it can only be concluded that no differences on the underlying scale are significant - it has not been shown that any second accents differ in their perceived prominences from any other.

4.4.5 Discussion

The clearest result from this experiment can be seen in Table 4.3, viz. that for about half of all the stimulus pairs, the second accented syllables were judged equal in prominence. This is corroborated by clustering of values on the putative underlying scale. As a result, we can conclude tentatively that there is some corroboration for the LDH and that we should reject the GDH; that the lack of any local declination has meant that second accents are judged to have about the same prominence, even when there is a prolonged interaccentual stretch allowing for an increased decline in any putative reference declination function.

However, there are some caveats which need to be attached to this interpretation of the results. Firstly, it has to be remembered that the task that subjects had to perform was comparison of the prominence of the second accents in the two sentences of a pair, which is not the same as comparing the prominence of the second accents relative to the prominence of the first accents in the two sentences. The earlier discussion of the hypotheses concentrated on this latter comparison, and, though it is true that a

comparison of the absolute second-accent prominences should give the same qualitative results under each of the hypotheses, removal of the effect of the prominence of the first accent may have attenuated the results.

Secondly, there is one possible condition not considered before conducting the experiment which would have been impossible to detect by means of it. This is the case in which the abstract declining baseline is of such a form that the pitch height of the second accent is the same in all six stimulus types, that is, regardless of the inferred duration of the utterance (as cued by the existence or otherwise of the interruption plus filled hesitation pause). This could be the case if the baseline was unaffected by forward-planning by the speaker, and happened to be at the same value at the point of the second accent in the case of the short utterances as in the long utterances. One way that this could be the case would be if the drop between the baseline at the point of the first accent peak and the baseline at the point of the second accent peak were the same in all six stimuli, and the baseline resulted from passive physiological activity by the speaker (or the inference of passive physiological behaviour by the hearer). Now this might be the case if the exponential curves in Figure 4.8 were tilted and added together to produce a narrowly varying sigmoid curve; for instance, if the abstract declination line were the result of the summed contribution of relaxation curves of cricothyroid muscle activity and declination curves of subglottal pressure decline (cf Collier 1987).

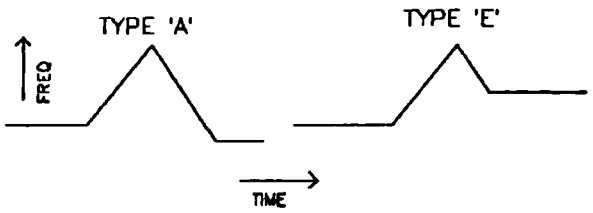
Thirdly, some who performed the test commented on how difficult it was to perform. This means that people may not have been able to adopt a consistent strategy, and in the absence of easily retrievable cues, may have tended to make judgments of equality in many cases, and several random judgments in others.

Fourthly, one or two subjects reported having adopted an analytical listening strategy, such as maintaining the first instance of the second accented syllable in echoic memory for comparison with the second when it occurred (though accounting for such strategies is a general problem in perceptual experimentation). Such a strategy necessarily precludes the direct operation of any declination function, be it local or global.

Fifthly, some subjects reported that for the most part, the second accented syllables sounded the same, and so they often had to make judgments of equality, although through their expectations that some stimuli should exhibit differences in the prominence of the second accented syllables, they forced themselves to make inequality judgments in some cases.

The fact that the second accented syllables sounded the same may well simply be the result of their being identical acoustic tokens, albeit in different contexts. The fact that for each stimulus pair, exactly the same [ja] sequence occurs four times may have resulted in habituation to the stimuli occurring among subjects, such that it was impossible to avoid the perception that second accented syllables had the same prominence. This observation leads to the conclusion that any similar such experiments should avoid those problems by making the repeated accented syllable match in terms of its prosodic parameters, but not be identical to the first instance of it.

Sixthly, another factor that may have arisen was that the stimuli may have appeared unnatural in the form of the intonation contours with which they were resynthesized. Study of 'Cursus Nederlandse Intonatie' (Collier and t'Hart 1978) revealed that some form of final lowering would always be expected in Dutch, if only because of the declination line 'poking its tail out'. Also, one of the Falling contours listed in that course, Type 'E', is an arrested fall, thus :



It may have been the case that the second accented syllables in stimuli 1 and 4 appeared to exhibit contours intermediate between natural instances of Type A Falls and Type E Falls. This ambiguity might also have tended to make subjects listen in an analytic fashion.

One last comment that needs to be made is that the ordering effect observed in the experimental results is perhaps the result of cross-stimulus declination, or similarly 'downstep'[16]. It could, on the other hand, be the result of an abstract declination function, but whether it were and could be quantified would require further investigation to determine. In this regard, the ordering effect might be an instance of the phenomenon of backward masking. How much of any 'declination effect' is backward masking would need to be explored, but is not done so in this thesis.

4.6 CONCLUSION

The results from this experiment give some support to the Local Declination Hypothesis, but do not go nearly far enough to disprove the Global Declination Hypothesis. This is largely because there is an inherent difficulty in testing for the existence or otherwise of an abstract phenomenon which in principle could take a number of forms and could be cued by a number of other factors, some physically observable, some in principle unobservable. Falling into the latter category are the expectations of the speaker regarding what normally occurs prosodically within an utterance. It is impossible to monitor what these are, to know whether they are operative in the perception of a particular stimulus, and to control for them, and they may change from utterance to utterance. For instance, it is impossible to say whether a subject in the kind of listening experiments discussed and performed here expects there to be final lowering and compensates for it in scaling the prominence of an accent even in the absence of its occurrence. It is only possible to look at the results and make some post-hoc suggestions about the existence or otherwise of the expectation.

The only way of seeing what such expectations concerning declination exist in the act of the perception of intonation and coming up with some appropriate quantitative analysis of the 'abstract' component of declination is to see how much of the phenomenon can be accounted for by a model which looks only at the relationship between the scaling of prominence and the physically observable variation in the F0 signal. This can be done by investigating the Local Declination Hypothesis directly, and it is this which is done in the following chapter.

---

[16] As could also be the case in some of Hermes and van Gestel's (1991) stimuli.

4.6 APPENDIX 1 - CRITIQUE OF LEROY'S ANALYSIS

Leroy's analysis of the results she presents in Leroy (1984) can be criticised on three counts.

POINT 1

Firstly, some of the z-score data values have been wrongly recorded. In her condition 'S', her graph of weighted average z-scores has a value of something like -2.1 for the lowest value of P2-P1 (being 8.22 ST), i.e. stimulus S-4. It is not clear where this value has come from. The correct value in this position is
-1.36. This is derived in the following way :

Count of H responses for S-4 = 3
H-response/Total responses for S-4 = 3/84
Count of E responses for S-4 = 12
(H+E responses)/Total responses for S-4 = (3+12)/84
z-score for H = z-score for 3/84 - 0.5 = z-score for -0.464 = -1.8
z-score for H+E = z-score for 15/84 - 0.5 = z-score for -0.321 = -0.92
Mean z-score for H and H+E = (-1.8-0.92)/2 = -1.36

When the PSE is calculated (as the point on the X-axis corresponding to the zero-point on the Y-axis of the linear regression of the z-scores on the values (P2-P1)) with the z-score value of -2.1 for stimulus S-4, the value obtained is -2.08, as Leroy records it. When the correct z-score value of -1.36 is used, the PSE shifts left to a value of -2.43 (+/- 0.98).

Unfortunately, this doesn't enhance the significance of Leroy's results, as it might be expected to, because there also appears to be an error in the data values she has recorded for condition IS. This time the error seems to be a straightforward one of sign-reversal. The z-score for stimulus S0 (the stimulus for which, objectively, P2=P1) should be 0.505, and is derived in the same way as demonstrated above, from an H-response count of 18, and an E response count of 63. However, Leroy has recorded this as a value of -0.505, and this shows up as a point some way off the main trend of the linear regression in her Figure 10 (Leroy 1984, p.61). Again, when the PSE is calculated with Leroy's recorded z-score of -0.505 for stimulus S0, it comes out as -0.32, as she recorded it. With the correct z-score of 0.505 for that stimulus, it comes out as -0.6 (+/- 0.22).

It should be noted that there is a similar sign-reversal error for the S0 z-score appearing in the graph for her condition 'M' (Leroy 1984, p.57), but this time the correct value (of 0.305) appears to have been used in the calculations leading to a PSE of -1.03.

POINT 2

Secondly, as Leroy herself noted, one way of apportioning the 'E responses' amongst the H and L responses is to divide them equally between them; that is, prior to calculation of the z-scores, the E responses could be partitioned half to the H-responses and half to the L-responses. We can call this the prior partition method. The values computed by prior partition are somewhat different from the values computed by Leroy's method of partition. The PSE values (with Standard Error of measurement values in brackets) and t-tests for significance as recorded by Leroy, the values they should have been according to the adjustments required by the first point above, and the values they would have been if the prior partition method had been used are all summarised in Table 4.A.1 overleaf.

	Leroy's data	Leroy's data (adjusted)	By prior partition
Cond.	PSE	PSE	PSE
N	-1.81 (+/- 0.89)	-1.85 (+/- 0.87)	-1.98 (+/- 0.82)
S	-2.08 (+/- 0.55)	-2.43 (+/- 0.98)	-2.44 (+/- 1.1)
M	-1.03 (+/- 0.41)	-1.03 (+/- 0.45)	-1.21 (+/- 0.44)
IS	-0.32 (+/- 0.58)	-0.6 (+/- 0.22)	-0.54 (+/- 0.38)
t-PERM	t-test result	t-test result	t-test result
S+N(11)	t=0.26,p>>0.1	t=0.44,p>>0.1	t=0.34,p>>0.1
M+N(11)	t=0.8,p>>0.1	t=0.84,p>>0.1	t=0.82,p>>0.1
M+S(10)	t=1.53,0.1>p>0.05	t=1.3,p~=0.1	t=1.04,p> 0.1
IS+S(11)	t=2.2,p<0.25	t=1.82,0.05>p>.025	t=1.63,0.1>p>.05
IS+N(12)	t=1.4,0.1>p>0.05	t=1.39,0.1>p>0.05	t=1.59,0.1>p>0.05
IS+M(11)	t=1.0,p>0.1	t=0.86,p>>0.1	t=1.15,p>0.1

(Figures in brackets after t-PERM indicate the degrees of freedom).

Table 4.A.1 – PSE values and t-tests for significance for the four conditions in Experiment 1 of Leroy 1984, computed (a) from Leroy's recorded data, (b) (a), adjusted, and (c) by the prior partition method.

These results indicate that no combination of unpooled conditions with the adjusted data is significant at the $p<0.025$ level, although condition IS is significantly different from S at the $p<0.05$ level. When the values are computed using the prior partition method, no condition reaches significance even at the $p<0.05$ level.

POINT 3

Thirdly, it appears that Leroy is not justified in claiming that the pooled conditions (IS + M) and (N + S) have significantly different PSE's at the $p=0.05$ level. Calculations from the adjusted data for that condition yield a t-test result of $t=1.178$ ($df=24$, $p > 0.1$). However, it is appropriate to note that in working through the calculations, results similar to Leroy's could be achieved if the results of intermediate calculations were rounded down to 2

decimal places. This loss of intermediate precision can clearly have quite marked effects on final results. In the particular case in point, this is the outcome: the t-test for the significance of the difference between two PSE's, using the formula that Leroy cites (Leroy 1984, p.96) is very sensitive to variation in the value of the Standard Error of Measurement (s.e.m.) (the formula is as follows:

$$t = abs(PSE1 - PSE2) / sqrt(s.e.m.1^2 + s.e.m.2^2)).$$

In computing the values for the t-test just reported (t=1.178) the value of s.e.m.1 (i.e. for (M + IS)), rounded to three decimal places, was 0.864, and for s.e.m.2 (i.e. for (N+S)), similarly rounded, was 0.898. If the intermediate calculations for these variables are rounded to two decimal places after the calculation of each term in the formulae for the terms of the formula for s.e.m. provided by Leroy, then both values are calculated as 0.60 (to two decimal places). If these values are used in the computation for the t-test result, the result is t=1.73, p~= 0.05, i.e. significance at the p=0.05 level is also achieved, misleadingly, for the adjusted data.

Even this result is not achieved if the data from the prior partition method are pooled. In this case, with 5-decimal place precision maintained in computation, the result for the t-test for the significance of the difference of the PSE's for condition (M+IS) against (N+S) (with s.e.m.1 = 0.902 and s.e.m.2 = 0.921, to 3.d.p) is t= 0.866, df=24, p >> 0.1. With only 2-decimal place precision maintained in computation, the result (with s.e.m.1 = 0.68, s.e.m.2 = 0.69), is t=1.15,p>0.1.

4.7 APPENDIX 2 - LISTENING EXPERIMENT INSTRUCTIONS

The following instructions were given to subjects in the listening experiment[17]. A translation into English follows the Dutch instructions.

Luisterexperiment: INSTRUCTIES

U gaat nu een aantal zinsparen horen die door een man zijn uitgesproken. In elke zin van een paar zijn twee woorden meer beklemtoond dan de andere.

Uw taak is het om de mate van beklemtoning van het tweede beklemtoonde woord in zin EEN te vergelijken met de mate van beklemtoning van het tweede beklemtoonde woord in zin TWEE.

In de zinnen wordt gesproken over het bereiden van limonade van een onbekende tropische vrucht, de 'pajaar'. Het kost geruime tijd om van een dergelijke vrucht sap te maken.

Bij het beoordelen van de mate van beklemtoning van de betreffende woorden moet van de onderstaande schaal worden gebruik gemaakt.

Hierbij is:

Z1	:	klemtoon van het tweede woord in zin EEN
Z2	:	klemtoon van het tweede woord in zin TWEE
>>>	:	VEEL sterker beklemtoond dan
>>	:	STERKER beklemtoond dan
>	:	IETS beklemtoonder dan
=	:	GELIJK beklemtoond als

Z1 >>> Z2 Z1 >> Z2 Z1 > Z2 Z1 = Z2 Z2 > Z1 Z2 >> Z1 Z2 >>> Z1

Deze schaal vindt U voor elk zinspaar op de volgende pagina's. U moet omcirkelen wat U voor een bepaald zinspaar van toepassing acht.

Wanneer U de band start hoort U eerst 5 random gekozen zinnen van het type dat U in dit experiment gaat beoordelen.

Daarna hoort een piepje, en begint het experiment.

U hoort 30 zinsparen, opgedeeld in blokken van 10.
Tussen elke zin van een paar is een pauze van 1 seconde, en tussen de paren zijn pauzes van 4 seconden; U kunt dan Uw antwoord geven.

Tussen de blokken is een pauze van 10 seconden; U kunt dan de bladzijde van het scoreboekje omslaan.

Even een voorbeeld:

U hoort:
'van pajaren maken we na jaren..eehhh'

---

[17]Thanks are due to Toni Rietveld for his translation into Dutch and amelioration of some instructions in English that I originally drafted.

'van pajaren maken we in Leeuwarden na jaren..eehhh'

Als U de klemtoon op 'jaren' in zin 1 enigzins sterker vindt dan die op 'jaren' in zin 2, omcirkelt U 'Z1 > Z2'.

NB. Het gaat hier om het woordje 'jaren', en NIET om het woordje 'pajaren'.

Hartelijke bedankt voor Uw medewerking.

TRANSLATION

Listening Experiment: INSTRUCTIONS

This time you're going to hear a number of sentence-pairs spoken by a man. In each sentence of the pair two words are more accentuated than the others.

Your task is to compare the degree of prominence on the second accentuated word in sentence ONE with the degree of prominence on the second accentuated word in sentence TWO.

The sentences are about the production of lemonade from an unknown tropical fruit, the 'pajaar'. It takes a considerable time for juice to be produced from such a fruit.

In judging the amount of prominence on the relevant words, use should be made of the scale below.

In the following, the meanings of symbols are :

Z1	:	prominence of the second word in sentence ONE
Z2	:	prominence of the second word in sentence TWO
>>>	:	MUCH more prominent than
>>	:	MORE prominent than
>	:	SOMEWHAT more prominent than
=	:	EQUALLY prominent as

Z1 >>> Z2 Z1 >> Z2 Z1 > Z2 Z1 = Z2 Z2 > Z1 Z2 >> Z1 Z2 >>> Z1

You'll find this scale against each sentence-pair on the following pages. You should circle that which you consider applicable to a particular sentence-pair.

When you start the tape you'll hear in the first place 5 randomly selected sentences of the type that you're going to judge in the experiment.

Then you'll hear a beep, and the experiment will begin.

You'll hear 30 sentence pairs, divided into blocks of ten.
Between each sentence in a pair there is a 1 second pause, and between the pairs there are pauses of 4 seconds, during which you can give your answer.

Between the blocks is a pause of 10 seconds, when you can turn the page of the scorepad.

An example:

You hear:
'van pajaren maken we na jaren..eehhh;
'van pajaren maken we in Leeuwarden na jaren..eehhh'

If you find the prominence on 'jaren' in sentence 1 somewhat stronger than that on 'jaren' in sentence 2, you circle 'Z1 > Z2'.

N.B. The word of interest here is 'jaren', and NOT 'pajaren'.

Many thanks for your cooperation.

CHAPTER 5
THE FORM OF DECLINATION

5.1 INTRODUCTION

In the previous chapter, two different hypotheses concerning the nature of a putative declination function were discussed. The first of these, the Global Declination Hypothesis was tested directly, by seeing whether stimuli which had no local declination in them (apart from a bit of final lowering in the case of two of the stimuli) elicited the declination effect. The results, although seeming to favour the opposing position expressed in the Local Declination Hypothesis, were ultimately seen to be inconclusive, largely because of difficulties in assessing the type of abstract cues listeners might be using in the perceptual test, allied to difficulties in estimating what form of global (abstract) declination function listeners may have used in making prominence judgments.

In this chapter, the Local Declination Hypothesis will be examined more directly. A general model is developed of individual accent prominence scaling, derived from the additive combination of local peak height factors and local contextual factors (which can, on occasion, be local declination factors). The emphasis is on the derivation of a perceptual prominence rating on the basis of intonational features which are physically present in the F0 signal[1]. It is shown that this can be used as the basis of a model of the production of intonation, in which the F0 values on consecutive accent peaks of a particular target prominence are predicted on the basis of the values of earlier peaks and the surrounding unaccented material. The approach allows the processes of perception and production of intonation to be considered at the same time, and some predictions to be made about the process of the production of intonation (in which production and perception are combined) in readiness for a production experiment reported on in Chapter 6. At the same time, it is made explicit which aspects of the declination effect are due to local declination and which to downstep, which function takes on more global characteristics within the model.

---

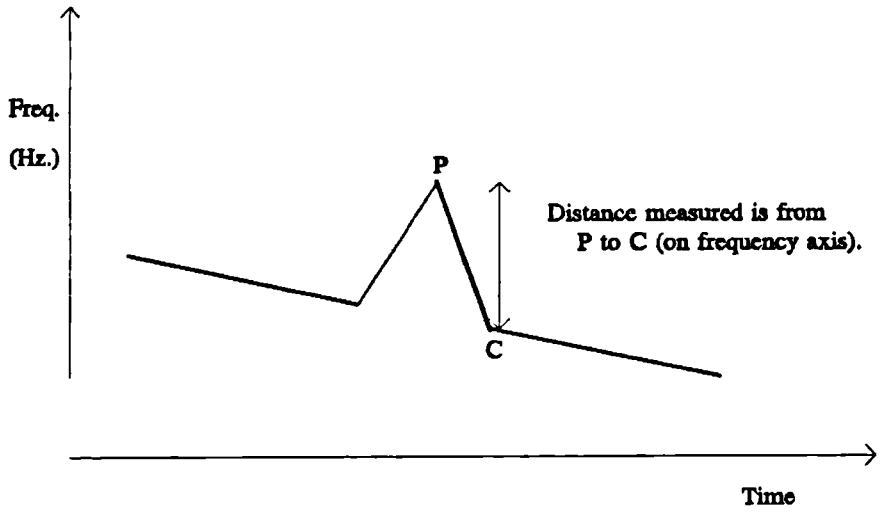
[1] As has been mentioned previously (footnote 2 in Chapter 2), this can be interpreted as a criterion which minimises the degree of abstraction in the perceptual cues which are used in the rating of prominence, since there could be said always to be some degree of abstraction involved in the process of perceiving intonation.

5.2 THE SCALING OF INDIVIDUAL ACCENT PROMINENCE

In the previous chapter, some experiments by Terken were discussed in which two competing hypotheses were tested, the CHANGE hypothesis and the MAX hypothesis. These proposed, respectively, that the prominence of an accent relative to its neighbour is a function of their respective distances from a putative (declining) baseline, or a function of the respective F0 values on their peaks (their F0 maxima). We can interpret the use of the word 'relative' used in these hypotheses as meaning that the prominence of any individual accent is a unitless quantity which is given an interpretation only in expressions of the form 'Accent A is x times more/less prominent than Accent B'. The first task in this section is to examine the concept of prominence as it relates to individual accents, partly to see whether it can be quantified.

5.2.1 Prominence as a function of peak and baseline

The first step in that process is to consider what salient points in the intonation contour could be used as a basis for a prominence function, given the requirement that such points have some physical manifestation in the F0 contour. Terken's CHANGE hypothesis expressed the idea that the prominence of an accent depends on the distance of its peak from a baseline. This distance is taken by Terken to be the vertical distance, in other words to a point on the baseline which doesn't actually appear in the contour (see Chapter 4, Fig. 4.5). Since the current approach aims to reduce to a minimum the amount of abstraction involved in the determination of prominence, however, the appropriate distance to measure, for a peaked accent such as that used in the experiments discussed in Chapter 4, is from the accent peak to the next point which appears on the baseline, thus:



Here, the frequency axis is in units of Hz., and the question immediately arises whether a scale more appropriate to changes in frequency, such as the semitone scale or the ERB scale, should not be used. The answer to that question lies in consideration of the basic function which ought to be used to represent the difference in frequency between peak and baseline, or between any point on the frequency axis. Terken tests a hypothesis (CHANGE) which uses frequency difference as the basic function, but it is worth considering whether it isn't the use of that function itself which raises the problems about choice of an appropriate scale for intonation. Another primitive function which expresses the same relationship between higher and lower points (the higher the point P in the above figure, the higher the value of fn(P,C), ceteris paribus) is that of division. Then, for the above example, the ratio of the two points P and C can be taken as the function (see F1 below) expressing the first approximation to the prominence of the peaked accent in the figure.

F1: $p(P,C) = P/C$

This approach has one immediate advantage, which is that the same relationship between P and C corresponds to the same value of prominence wherever one is on the frequency scale. For instance, if P is 150 Hz and C is 75 Hz, the value of p(P,C) is 2, and the same is true if P is 204 Hz and C

is 102Hz. Similarly, if P is 300Hz and C is 100Hz, p(P,C) is 3, and likewise if P is 240Hz and C 80Hz.

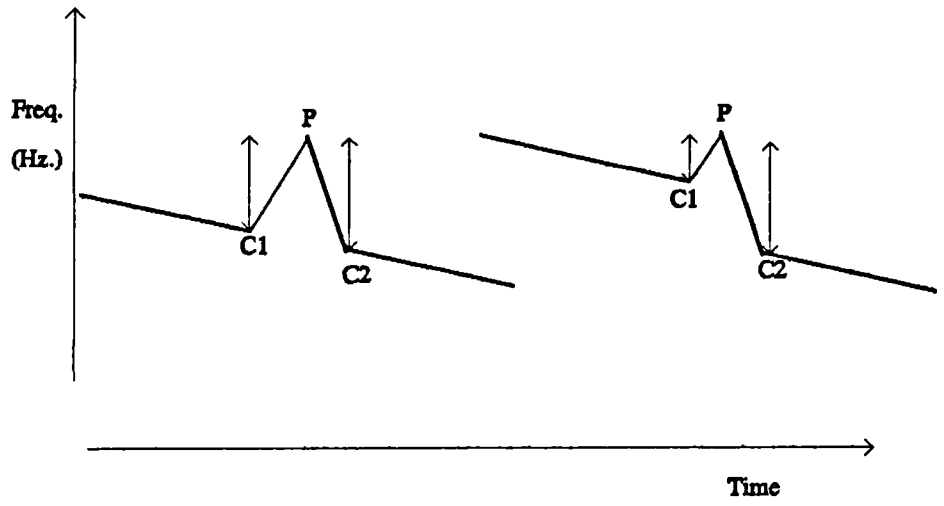
It also has one apparent disadvantage, which is that it allows for prominence values to be expressed for frequency values which couldn't occur within a particular speaker's range. Thus, for any speaker imaginable, the ratio of 50/10 can't be derived (few speakers have vocal folds which vibrate at 50 Hz, and none have them which vibrate regularly at 10Hz in normal voice), yet that ratio corresponds to a meaningful prominence value of 5 using function F1. This is in contrast to the function Pierrehumbert (1980) uses to scale the salient points in her intonation contours (and associated prominence values, as discussed in Chapters 2 and 3), which has an implicit absolute baseline specific to an individual speaker, and which also allows the same value of prominence to be expressed for the same ratio of frequency expressions wherever in the frequency range they occur:

F_p: $p(P,C) = (P-C)/C$

The reason that this function has not been adopted is that it is not needed for the operations that will be detailed later, because specific constraints on the operations prevent unnatural values being assigned to P and C; thus, the simpler form expressed in F1 is used as the basic prominence function. F1 also has the advantage of entailing the utilisation of a simpler speaker-normalisation function by hearers of another's speech, without requiring inference or observation as to the speaker's absolute baseline.

5.2.2 Prominence as a function of peak and preceding baseline

It is suggested here that the prominence of a peaked accent, such as that discussed in the previous subsection, depends not only on the ratio between the peak F0 and the next audible point on the baseline, but also on the last previously audible point on the baseline. According to this hypothesis, in the following figure, the prominence of the leftmost accent would be greater than that of the rightmost.



For the time being, it is suggested that the contribution to the prominence of the accent of the points P and C1 is the same as that of points P and C2. If the prominence of the accent were taken as the mean of the ratios P/C1 and P/C2, then the prominence of the leftmost accent above would be the same as the ratio between P and the point vertically beneath P on the baseline projected between points C1 and C2. In this case, then, it would be possible to see how part of an abstract declination function could be built up from the 'concrete' composite function expressed in F2:

F2: p(P,C1,C2) = (P/C1+P/C2)/2

That that is unlikely to be the whole story of how such a putative abstract function might be constructed can be seen from the rightmost contour, in which the point vertically beneath the peak P on the most likely projected one-piece baseline (which includes the declining stretch to the right of the accent peak) is lower than the point having a value according to function F2. In fact, for reasons that become apparent later, the prominence of the accents is taken as the sum of the ratios P/C1 and P/C2, as expressed in F3:

F3: p(P,C1,C2) = P/C1 + P/C2

5.2.3 Prominence as a function of peak duration

The representation of peaked accents in the above figures is highly stylised. The rapid change in direction in the intonation contour at the position P is unrealistic; the change in direction is typically more gradual. This is reflected in the fact that in standard Dutch stylisations within the Eindhoven School, a plateau of duration of the order of 30ms. is inserted between the rise and fall of such accent contours (these are called 'pointed hat' accents). Such an adjustment to the contour has the effect of boosting the percept of the peak pitch (in fact, bringing it into existence, since, as 't Hart et al. (1990) note, a duration of c.30ms is the minimum required for the pitch of a speech-like stimulus to be perceived). As a consequence, the prominence of the accent is increased.

This relationship between peak duration and prominence needs to be expressed as some factor in the function being developed to define the prominence of an accent. It is not done so here, firstly because it is premature to make a quantitative estimate of the contribution of such a plateau, since it involves consideration of the contribution to prominence of level stretches of F0 (see below), and secondly for the following reason. The use of straight line contours to represent the intonation patterns that has been adopted in this and the previous chapter reflects the desire to represent the contours as straightforwardly as possible. There exists an equally straightforward representation which involves the use of parabolic interpolation between target points (see Hirst 1983). As it stands, the prominence function is equivocal about what sort of interpolation takes place between the points C1, P and C2, but parabolic interpolation has the advantage of flattening at the peak in something like the way required for realism and elicitation of the correct prominence-pitch ratio for an accent whose peak has a particular nominal F0 value. So, it would be possible to stipulate that the prominence function being developed should be understood to be used in conjunction with a model employing parabolic interpolation between the salient points incorporated within it. However, the model developed in this chapter is designed not to be constrained by methods of interpolation. Strictly speaking, it should be able to model the prominence of the components of any contour, be it straight-line stylised synthetic, with parabolic interpolation, or a raw one. Thus, such a stipulation is not made here, and the question of the contribution of peak duration to prominence is

left open for the time being. However, considerations relating to that question and the form of interpolation between salient points in a contour will arise regularly through the course of the coming discussion.

5.2.4 Prominence as a function of slope

As the time varies between points C1 and P and P and C2 (we may call the stretches between these and similar points contour elements) so does the slope of the rise and fall comprising the peaked accent, all other things being equal. In terms of capacity for accentuation, it is fairly clear that a fall of say, an octave, will be less prominent if extended over ten seconds than if extended over a fifth of a second. This fact is probably a result of a combination of two factors: the greater textual material that can be uttered in ten seconds compared with a fifth of a second span, which means that no individual textual item other than the first will be highlighted by use of such a contour; and an inherent pitch prominence that arises from the rapidity of the pitch movement over 200ms compared with that over ten seconds. In respect of this second factor, the question arises what function best describes the variation of prominence with slope of contour element.

The Eindhoven school half-answers this question; they posit in the inventory of primitives both a fixed slope fall and rise and a fall and rise of variable slope (see Chapter 2). The fixed slope has a value of one octave per 160ms, and it is implied that the prominence of any observed peaked accent depends on the deviation of its slope from this canonical slope[2], at which prominence is maximal.

The duration of 160ms over which the steep fall and rise of the Eindhoven School inventory are defined is an important constant. It is close to the value of 0.167 secs, or one sixth of a second, which can be taken to be the typical duration of a syllable at an average rate of production. The frequency 6Hz is within the range of frequncies of theta rhythm, slow wave activity in certain parts of the brain - particularly the hippocampal formation and reticular formation (see, for instance, Vinogradova 1976, Bland 1986, Vertes

[2] That this slope is at least of utility in describing the intonation contours of English has been jointly confirmed by the author during research developing an algorithm for intonation in text-to-speech synthesis (Johnson 1990, House and Johnson 1987).

1982) - to which has been attributed the function of gating of information to and from short-term memory. It is not unreasonable to suggest that the transmission of information in intonation is linked to the timing of an operative syllable clock (this idea being a well known one in certain theories of speech perception, where such a clock could be considered linkable to so-called P-centres in the speech stream (e.g. Marcus 1976)). To a first approximation, the period of this clock can be taken to be one sixth of a second. There is, however, no reason why it should not vary within the bounds of syllable rate observed in speech (although at high syllable rates informal observations of the course of rises and falls during fast speech suggest that the upper bound for the rate to which the timing of the start points of rises and end points of falls can be linked should be less than the maximum syllable rate in continuous speech).

This does not mean that accentual movements are fixed to the current syllable rate, but are linked to it; exactly how is not discussed here, though some relevant comments are made in section 5.2.7 below. But it is suggested here that the pitch prominence of accentual movements is measured by reference to a canonical duration. The value of 0.167 seconds is taken to be that duration (see also section 5.2.6 below). This is not just for heuristic reasons - apart from it being within the range of theta rhythm, it happens to be a multiple of 0.033 seconds, which is about the minimal duration for pitch perception of speech-like stimuli and Amplitude Modulated (AM) tones, as already mentioned in section 5.2.3. The reciprocal of that value, 33Hz, is also an approximate lower bound for a fundamental frequency eliciting the sensation of pitch. That frequency, or an approximation to it, has been suggested as a 'sampling rate' for any of the sensory modalities (Poeppel and Logothetis, 1986).

How, then, does the prominence of an F0 excursion of a particular extent in frequency vary as the duration of that excursion varies? For short excursions (less than 0.167 seconds), there are two basic possibilities: firstly, one could expect that as the duration decreases from the canonical duration, the prominence increases, the prominence being a direct function of (absolute) slope. Secondly, one could expect that as the duration decreases from that point, so does the prominence decrease, the prominence being temporally integrated below it. For longer excursions, again two

possibilities can be identified: firstly, it could be that there is a fairly rapid reduction in prominence beyond the canonical duration, to an asymptote of zero prominence; secondly, one could expect that as the duration increases above the canonical, so the prominence remains more or less the same, until perhaps a very long time later, when it gradually reduces to nothing. There are intermediate possibilities, of course, but these are not addressed for the purposes of exposition.

For short excursions, it is not really feasible to maintain the first possibility, that prominence is a direct function of (absolute) slope. As the duration of the excursion tends towards zero, so does the information transfer - the prominence can't increase in these circumstances. However, it could be that there is an initial increase in prominence as a function of slope as the duration decreases below 0.167 seconds, followed by a decrease as the duration approaches zero. This possibility is formalised in the first of four models of pitch prominence variation with duration, along with the first possibility for longer excursions. The equations for these models are introduced in section 5.2.6 ; their behaviour for an octave fall or rise is depicted in Figures 5.1 - 5.6. In each case, the main function (which appears as a solid line) is the product of two subsidiary functions: the frequency factor and the time factor (these appear as dashed lines in the figures, though the frequency factor is a constant in models 3 and 4, and so does not appear).

The four models are as follows :

Model 1: Exponential frequency factor * exponential time factor
Model 2: Exponential frequency factor * exponential two-piece time factor (as in Model 1 up to 0.167 seconds, then following a very slow sigmoid decay curve).
Model 3: Constant frequency ratio factor * exponential time factor
Model 4: Constant frequency ratio factor * exponential two-piece slow-decay time factor.

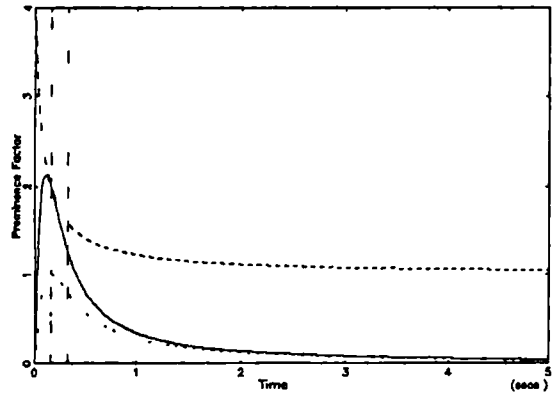


Figure 5.1 Pitch prominence factor variation with time - Model 1. Lower dashed line: time factor; upper: frequency factor; solid line: composite function. Vertical lines @ 0.167 and 0.334 secs.

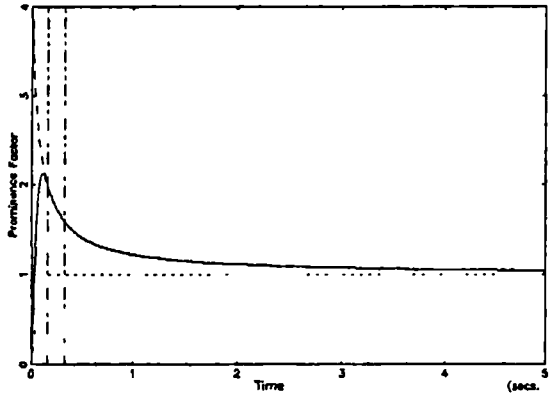


Figure 5.3 Pitch prominence factor variation with time - Model 2. Interpretation is as for Figure 5.1.

In these figures, the lower dashed figure has the same shape as the composite function would have for level contour elements, because the frequency factor for these always has a value of 1 (see section 5.2.6 for further discussion). In Figure 5.4, the function is extended over 40 seconds to show the temporal domain of the variation of the second part of the two-piece time factor function. It is not suggested that the course of this function is often relevant to the prominence of unary contour elements over such a time-span: the function acts as a constant for short to medium duration F0 excursions.

These functions are estimates of the possible variation of pitch prominence with slope for a fixed drop contour element[3]. What still remains is the requirement for some means of identifying points C1 and C2 (as in Fig 4.2), here in the environment of a peaked accent within the intonation contour[4]. In many cases, the manual identification of the points in a raw F0 contour is straightforward; there is a clear change from steep to shallow gradient both before and after the peaked accent. This is the case with the peaked accent on the word "rare" in Figure 2.13. In other cases, this is not

[3] They also estimate its variation with F0 drop for a fixed duration contour element.

[4] The question of the identification of such points in the case of step accents is addressed below.

possible, either because of the disruption caused by consonantal masking or influence on the F0 contour, or because the transition from accented syllable(s) to unaccented syllables is more gradual. The solution of this problem could be approached by the use of semi-automatic or automatic stylisation methods (for instance, Eindhoven school close-copy stylisation methods, as reported in 't Hart et al. 1990), the fitting of a quadratic spline curve to the F0 contour (Hirst 1983) or the fitting of a linear spline curve to it (Mead 1974). These methods are only part of the solution, however, since they require manual adjustment in some cases.

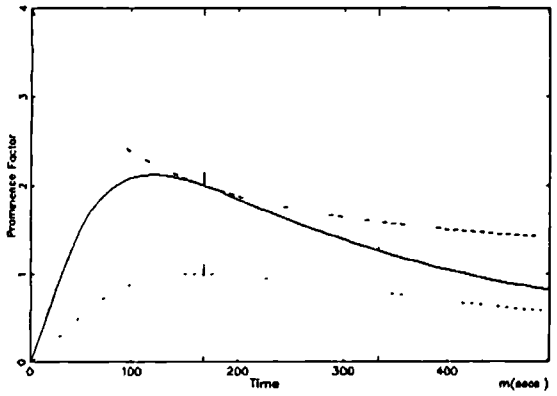


Figure 5.2 Pitch prominence factor variation with time - Model 1. Same function as in 5.1 displayed over half a second. Note composite function peaks before the 0.167 sec. point.

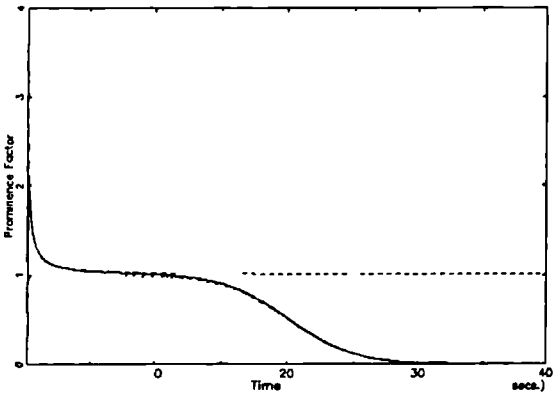


Figure 5.4 Pitch prominence factor variation with time - Model 2. Same function as in 5.3, displayed over 40 secs.

In any case, as is discussed below in the case of step accents and later in respect of one treatment of the head accents of falling heads (to use Crystal's (1969) and O'Connor and Arnold's (1973) terminology), there is a need to identify points in an F0 contour which don't necessarily correspond to some physically identifiable change in the F0 contour, but which occur at the 'gating' point in the contour relative to the syllable structure and operative syllable and contour rate. This being the case, the criteria for identifying points C1 and C2 in a contour which should be used in analysis of F0 contours in connection with the model being developed here will be as follows: if turning points in an F0 contour or stylisation thereof are clearly identifiable, corresponding to the expected pre-and post-peak positions of C1 and C2, and they are

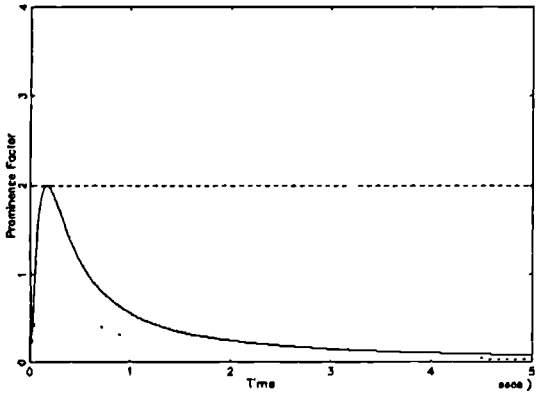


Figure 5.5 Pitch prominence factor variation with time - Model 3. Interpretation is as for Figure 5.1

within one syllable's distance in time from the peak P, where the duration of a syllable is determined empirically from the current speech sample, then those points should be marked as points C1 and C2. If no such points are identifiable, or if the peak can be identified at the left or right edge of a plateau, then points C1 and C2 should be marked at positions s seconds before and after point P (where s is, by default, 0.167 secs).

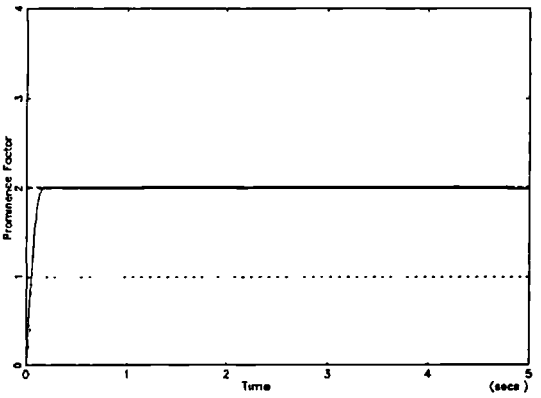


Figure 5.6 Pitch prominence factor variation with time - Model 4. Interpretation is as for previous figures.

The fixing of points C1 and C2 relative to the peak position is similar to the strategy that is adopted by autosegmental approaches to the analysis of intonation in the style of Pierrehumbert (1980), in which leading tones (e.g. the L in L+H*) and trailing tones (e.g. the H in L*+H; see Chapter 3 for fuller discussion) are taken to be at a fairly strictly circumscribed distance from the starred tone. The approach here differs in the requirement to identify two points rather than one relative to the peak (or trough - see below). Further work is required to identify where those points should be in any arbitrary contour.

5.2.5 Step Accents

In this analysis, step accents are accents which don't contain a peak, but which contain a plateau. Strictly, to qualify as a step accent, the F0 movement after the plateau should be in the same direction as that before the

plateau. From these basic step accents, downward, upward and complex stepping sequences can be generated, as in Figure 5.7.

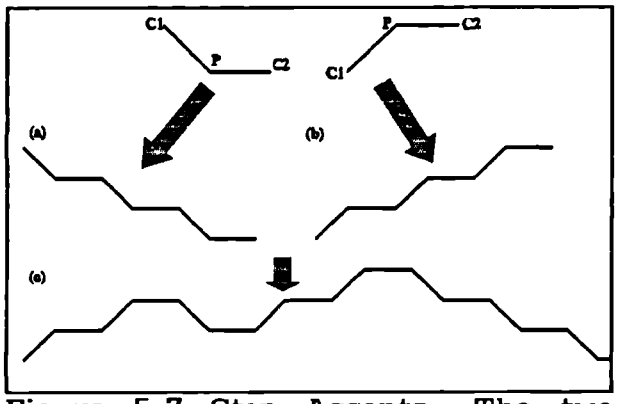


Figure 5.7 Step Accents. The two schemata at the top indicate the two basic types of step accent. Sequences (a), (b) and (c) can be derived by concatenation.

The important thing about the schemata for the present model is that the points C1, P and C2 are still identified with turning points in the contours at the top of the figure, even though they have a different relationship with each other from that in the peaked accent. In the 'stepped down to' step accent on the left, the right local maximum C1 is what was a local minimum in the peaked accent; the middle point P is a local minimum, the start of the step accent's plateau, whereas before it was the accent peak; and the left local minimum C2 is at the same height as P, whereas before it was at a point on a putative baseline. In the 'stepped up to' step accent on the right, the right local minimum C1 retains the relationship to P that it had in the peaked accent, but P is again at the start of a plateau, and C2 is again at the same height as P.

5.2.5.1 A general accent form

In the basic types of step accent shown, function F3 (p(P,C1,C2)=P/C1+P/C2) still serves to determine the prominence value. Now, if the principle of the mobility of points C1 and C2 is extended, so that either or both of the points may have an F0 value which is lower than, equal to or greater than that of P, in any accent, a general form for an accent is derived, as seen in Figure 5.8 below. However, this extension raises

problems in the derivation of a prominence value for any of the accent forms which can be so generated. In particular, allowing point C2 to be higher than point P allows rising[5] forms to be generated, which means that some consideration has to be given to what the relative prominence values of falls and rises should be, specifically where the pitch movement occurs over the same range of F0.

It has been noted in many places that, from a pragmatic point of view, a falling pitch movement represents dominance and a rising one represents recession or submission (e.g. Bolinger 1978, Ohala 1984). The question here is whether that means that a fall is more prominent than a rise. Given that the function being developed here is one of pitch prominence, it seems to be more cogent to argue that falls and rises of equivalent range have the same prominence. The function of pitch prominence is considered here to be a simpler one than that of a prominence mediated by pragmatic functions of dominance and recession, corresponding more to the activation of particular networks within the auditory pathway(s) and laryngeal motor pathway in the brain. (It is, though, an empirical question whether networks in the limbic system corresponding to the apperception of dominance and recession are more or less directly connected with those other networks. It is likely that the limbic system is involved in the generation of speech (cf. Juergens and Pratt, 1979a, 1979b, Davis and Zhang 1991).

If falls and rises over the same range of F0 are to have the same pitch prominence, then where the point C2 is higher than point P, and similarly where C1 is higher than P, they should swap roles in function F3. Thus, the pitch prominence function becomes F2' :

F2': p(P,C1,C2) = (max(P,C1)/min(P,C1) + max(P,C2)/min(P,C2))/2

---

[5] These are referred to as "rising" forms, where step accents which contain a rise (in the 'stepped up to' form) are not, because the rise in the former form commences at the point P. This point is more important in identifying the form of the accent because of its role in prominence as a function of alignment with syllable structure, as is discussed below in section 5.2.7 .

where max() and min() are functions defining the greater and lesser respectively of their two arguments. (Thus if P > C1 but P < C2, then p(P,C1,C2) = (P/C1 + C2/P)/2).

It has to be remembered that C1 is earlier in time than P, so that if P is greater than C1, the configuration before P is a rise. The reverse is the case if P is less than C1. This means that, both in the configuration before P and in the configuration after P, a fall and rise of the same duration contribute a commensurate degree of prominence. Level configurations before and after P that are equally long contribute equal prominence[6].

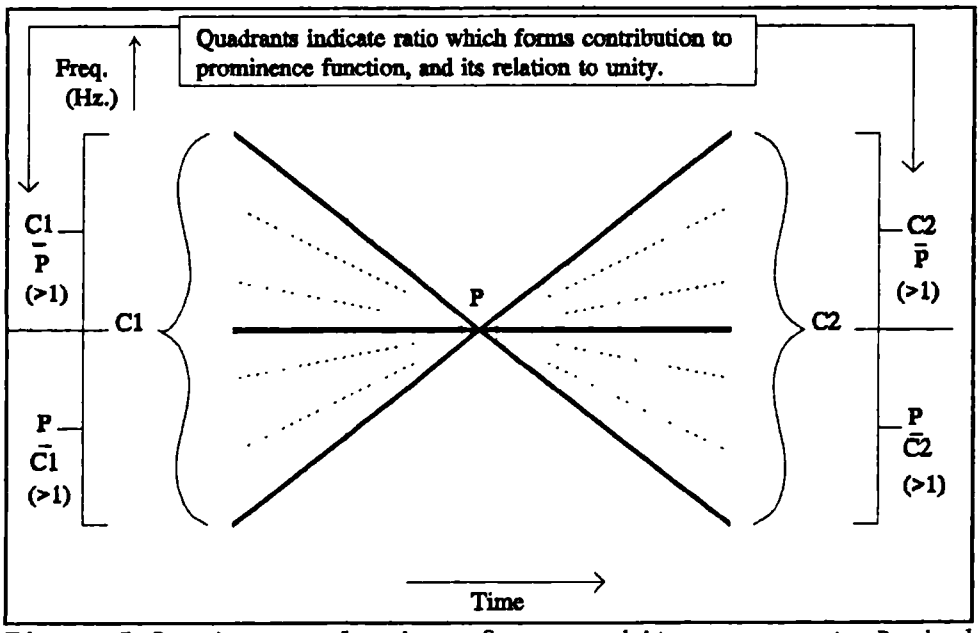


Figure 5.8 - A general schema for an arbitrary accent. Dashed lines simply indicate intermediate positions. There is in principle no restriction on the gradients of the lines in the model.

[6] It had been thought that in the configuration before P, a fall (i.e. where P1/C1 < 1) should be considered to have less prominence than a rise (where P1/C1 > 1), introducing an asymmetry in the prominence assigned to contour configurations before and after the point P. However, using that setting in the implemented model meant it didn't work as well (notably in downstep sequences). This is because the ratio is used as the base in the indicial frequency factor model introduced in section 5.2.4 and discussed in section 5.2.6. If this base is less than 1, an asymptotically increasing frequency factor function results, rather than an asymptotically decreasing one (for Models 1 and 2).

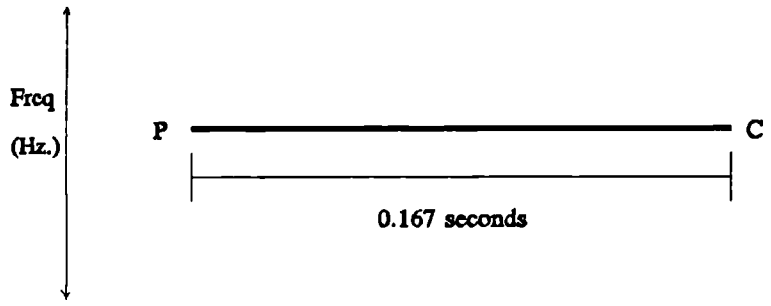
At this point it seems appropriate to bring up again the question of the flattening at the top of peaked accent peaks. Given the mobility of point C2, could the contour P to C2 not form the plateau in such a flattened peak? The answer is not certain. The degree of flattening in the peak is taken to depend on such matters as (1) physiological constraints (in particular, the degree of difficulty involved in the larynx effecting sudden changes in the rate of vocal fold vibration of the order required by a sharp accent peak), (2) the shape of the contour after points P and C2 and (3) segmental coarticulation effects. Only the second of these considerations is properly accounted for within this model - the first is beyond its scope and the third requires an account of compensation for phenomena 'known' to occur in speech by a speaker-hearer, which the current model is avoiding initially - and later discussion indicates the extent to which later parts of the contour interact with earlier parts. What can be said is that as a rule of thumb, peaked accent contours with a short plateau (say, less than 100ms) can be identified as having a single point P, with a shaping function around it, and those with a longer plateau can be considered to comprise an initial point P and a terminal point C2[7].

5.2.6 An initial approximation to a quantity for pitch prominence

If an arbitrary continuous[8] AM signal is considered for the moment, it is possible, from what has been put forward so far, to identify a specimen, whose value, on the basis of its shape and duration, can be used as a reference from which an initial approximation to the quantity pitch prominence can be given. That specimen is a level tone of 0.167 secs. duration, extending between two points P and C:

[7] In principle, it should be possible to specify any peak-plateau (above a minimal duration of, say, 30ms) as comprising a contour between points P and C2, given the common elements between the accent contour prominence functions and the unaccented contour prominence functions observed in the model (see below).

[8] That is, without breaks, silences or interspersed noise.



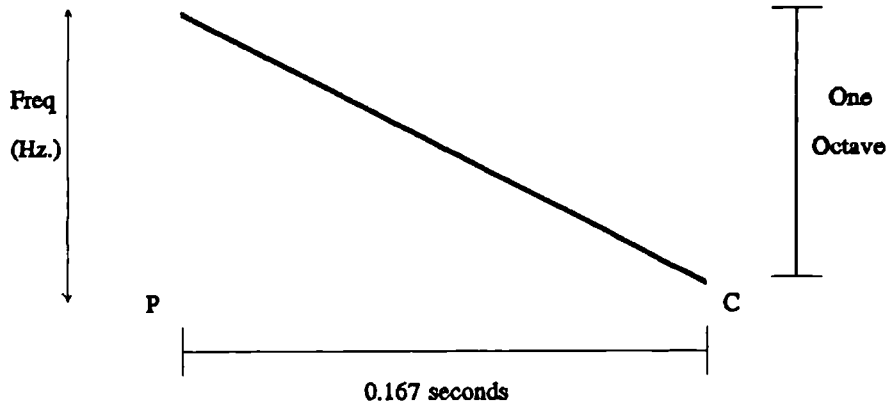
Note that the frequency axis arrow is double-headed. The intention is to show that pitch prominence does not depend on position within the pitch range, but on the relative positions of points such as P and C. The use of the word 'pitch', however, implies that some pitch should be determinable from the specimen; in particular, its fundamental frequency should be uniquely determinable and be within the range of human pitch perception. The application of the model of pitch prominence here requires only that an operative such range be specified within the frequency bounds of human speech intonation, say 33Hz to 1000Hz.

The ratio between point P and point C has a value of 1, regardless of the units with which P and C are expressed. If we take the units with which salient points are expressed to be Hz, since those are the units that are used as a basis for the ratios between points of differing magnitude in the frequency domain, then it is possible to define the ratio thus expressed as one pitch prominence unit (ppu):

D1 A continuous F0 contour extending over a period of 0.167 seconds with a constant F0 value has a prominence of one pitch prominence unit.

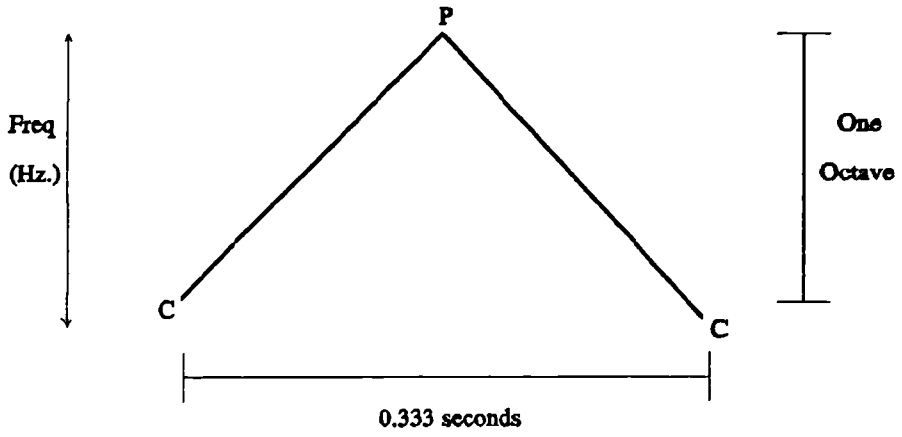
Since definition D1 makes reference both to the dimensions of time and frequency, it is natural to expect that variation of point C both in time and in frequency results in variation of pitch prominence. Taking the case of

variation in frequency first, we can see that a steep falling tone, falling over a range of one octave and lasting for the same duration, 0.167 secs., as the level tone of one ppu, has a value of two ppu :



Similarly, a steeper falling tone, falling over a range of 3xC Hz. in 0.167 seconds, has a pitch prominence of 3 ppu. Rises rising over the same frequency range and for the same duration as these falls have the same pitch prominence value.

When rises and falls occur together, as in the case of the peaked accent, the initial estimate of their pitch prominence presented here doesn't take into account any possible interaction, and the total pitch prominence is taken to be the sum of the parts. Thus, the pitch prominence of a peaked accent whose peak P has an



F0 value twice that of the local minima, C1 and C2, either side of it, as in the above figure, is 2+2 = 4ppu.

The effect of the variation of the temporal position of C relative to P on the pitch prominence of the P-C configuration has already been discussed in section 5.2.4. Here, some details of the functions introduced there can be given. Pitch prominence is a function not just of the pitch or pitch movement of a tone (or any stimulus) of arbitrary duration, but is modulated by such a tone's duration. At the extremes of the duration of a stimulus, it can be expected that this modulation effect is quite strong. An AM tone needs to be above about 30ms in duration for a pitch percept to be consistently associated with it (if the fundamental frequency is above about 100Hz, that is, if there are three or more pitch periods in the signal) ('t Hart et al. 1990); the pitch prominence of tones close to this duration is likely to be highly variable. It is also reasonable to expect that above a few seconds in duration, a habituation effect comes into operation (particularly for level tones), pitch prominence decaying as a function of the continued application of the stimulus.

These putative aspects of pitch prominence can be incorporated within an exponential function which maintains the pitch prominence value of a level tone of 0.167 seconds duration by being set to a value of 1 at that duration, but which decays to value of zero at zero duration, and has an asymptotic

decline to zero above 0.167 seconds. To ensure zero prominence at zero duration, and unity prominence at 0.167 seconds, the base of the function expression is made the ratio of the duration of the contour element to that canonical duration. So, if t(P) and t(C) are the start and end points in time of the contour element, the base of the expression is (t(C)-t(P))/0.167 . The exponent of the expression is constructed to provide satisfactory asymptotic behaviour without the addition of further empirical constants, and is set to 2*0.167/(0.167+t(C)-t(P)) - 1 . At zero duration, this exponent is 1, at 0.167 seconds it is zero, and above 0.167 seconds it tends towards -1, which means the whole expression in the following time factor function tends towards zero above that point.

The temporal modulatory behaviour of the pitch prominence function is thus incorporated within a function which can be called the time factor, T(P,C):

$$T(P,C) = ((t(C)-t(P))/0.167)^{2*0.167/(0.167+t(C)-t(P))-1}$$

This function appears as the lower dashed line in Figures 5.1-5.3, and is incorporated within Models 1 and 2 introduced in section 5.2.4.

A different hypothesis regarding the effect of increased duration on the pitch prominence of a tone has it that beyond the canonical duration, prominence is pretty much constant, decaying only after a long period of duration. This hypothesis is formalised in a two-piece model of the temporal modulatory behaviour of the pitch prominence function, the first piece identical to the time factor function just introduced, and valid when the duration of the contour element is less than or equal to 0.167 seconds, and the second piece a sigmoid curve with upper and lower asymptotes:

$$T'(P,C) = T(P,C) \text{ (for } (t(C)-t(P)) <= 0.167 \text{ seconds)}$$
$$= 1-1/(1+1.5^{20.0-t(C)-t(P+0.167)}) \text{ (for } (t(C)-t(P) > 0.167 \text{ seconds)}$$

This function appears as the lower dashed line in figures 5.4-5.6, and is incorporated within Models 3 and 4 introduced in section 5.2.4.

The contribution to pitch prominence of the F0 values of the start and end points of a contour element has already been encoded as the ratio between the

maximum value and the minimum value. This is one estimate of the function which can be called the frequency factor, F(P,C) :

$$F(P,C) = \max(P,C)/\min(P,C)$$

The product of the frequency and time factors can be taken as the composite function which expresses the main contribution to pitch prominence of an arbitrary contour element. For the simple frequency factor just expressed, this is the composite function which constitutes Models 3 and 4 introduced in section 5.2.4 . The view may be taken, however, that variation of pitch prominence is even more closely tied to the slope of a contour element than is implied even by the interaction between the frequency factor and the exponential time factor. It could be thought that in the absence of the influence of the time factor, higher slope corresponds to higher prominence. This hypothesis is formalised in an alternative exponential frequency factor, which uses the simple frequency factor as a base, uses the exponent of the time factor (plus 1), and is constructed to have a value the same as the simple frequency factor at the canonical duration (e.g. 2 for an octave fall or rise) :

$$F'(P,C) = (\max(P,C)/\min(P,C))^{2*0.167/(0.167+t(C)-t(P))}$$

This function is used as the frequency factor in Models 1 and 2 introduced in section 5.2.4, and appears as the upper dashed line in Figures 5.1-5.3 [9].

The effect of the interaction of this exponential frequency factor with the exponential time factor is for steepness of slope to have more influence on the pitch prominence factor. The maximum of the composite function is at a duration somewhat less than 0.167 seconds (see Figure 5.2; the greater the basic frequency ratio, the closer the position of that maximum is to zero, though it remains close to 0.167 seconds for meaningful frequency ratios). Also, beyond the canonical duration, the composite function decreases more rapidly towards zero. To preempt later discussion a little, such a function is appropriate for a model in which accentuation is more clearly demarcated

---

[9] Having included all the constituent factors of Models 1 to 4 in the discussion here, it is now appropriate to point out that the units on the Y-axis in Figures 5.1-5.6 should be pitch prominence units (ppu).

from a contextual function (such as a local declination function), and is done so on the basis of slope of the contour element.

The composite function is the next approximation to the pitch prominence function. There are four of these, corresponding to the four models of the variation of pitch prominence contribution with time:

$F4_{M1}$: p(P,C) = F'(P,C) * T(P,C)
$F4_{M2}$: p(P,C) = F'(P,C) * T'(P,C)
$F4_{M3}$: p(P,C) = F(P,C) * T(P,C)
$F4_{M4}$: p(P,C) = F(P,C) * T'(P,C)

An example of the different predictions the different models make of the variation with time of pitch prominence is as follows: For an octave fall, all models predict a value of 2ppu at 0.167 seconds' duration. At 0.334 seconds, Model 1 predicts a value of 1.26ppu, Model 2 predicts 1.59ppu, Model 3 predicts 1.59ppu and Model 4 predicts 2ppu. At 0.501 seconds, Model 1 predicts a value of 0.82ppu, Model 2 predicts 1.41ppu, Model 3 predicts 1.15ppu and Model 4 predicts 2ppu.

These functions are applicable regardless of the denotations of points P and C; in the accent diagrams within this subsection, points P and C can be reversed within either factor, the functions T(P,C) and T(P,C)', and F(P,C) and F(P,C)' remaining valid. Therefore, the functions are applicable independently to the rise and fall parts of a peaked accent, and the set of pitch prominence functions for that configuration becomes:

$F5_{M1}$: p(P,C1,C2) = F'(C1,P) * T(C1,P)
+ F'(P,C2) * T(P,C2)
$F5_{M2}$: p(P,C1,C2) = F'(C1,P) * T'(C1,P)
+ F'(P,C2) * T'(P,C2)
$F5_{M3}$: p(P,C1,C2) = F(C1,P) * T(C1,P)
+ F(P,C2) * T(P,C2)
$F5_{M4}$: p(P,C1,C2) = F(C1,P) * T'(C1,P)
+ F(P,C2) * T'(P,C2)

The following is the full expansion of $F5_{M1}$:

$$F5_{M1}: p(P1,C1,C2) = (\max(C1,P)/\min(C1,P))^{2*0.167/(0.167+t(P)-t(C1))} * ((t(P)-t(C1))/0.167)^{2*0.167/(0.167+t(P)-t(C1))-1} + (\max(P,C2)/\min(P,C2))^{2*0.167/(0.167+t(C2)-t(P))} * ((t(C2)-t(P))/0.167)^{2*0.167/(0.167+t(C2)-t(P))-1}$$

Thus, an initial approximation has been given to a quantity for pitch prominence as a function of configurations for AM tones, with quantitative estimates of the variation in pitch prominence as a function of variation of salient points within the configuration in both the frequency and time domains. The expressions defining this quantity are used in later sections as further development of a model of individual accent prominence for accents occurring within the speech stream, and it is in that context that their validity is tested by incorporation within a working computer model which predicts the height of successive accent peaks in an intonation contour.

5.2.7 Prominence as a function of accent alignment to syllable structure

Having looked at the effect on pitch prominence of the positions of points in an accent configuration relative to each other, it is time to include consideration of the effect of positions of points relative to the segmental material in speech. This can be done by recourse to the elements of syllable structure.

In any language, an arbitrary syllable has a mandatory component (the peak), which consists of a vowel or vocalic consonant, with an additional selection of at least one from two optional components, one preceding the peak, called the onset, and one following it, called the coda. Both these latter components comprise consonantal material. This principle of delineation of the elements of the syllable is fairly well accepted[10]; what is still unclear are the principles for syllabification, that is, where syllable boundaries should be placed in languages which have both onset and coda, and what the processes are of resyllabification (the realignment of part or all of an onset

[10]although the grouping of these units varies from theory to theory. Most commonly, the peak and coda form a unit together called the rhyme (Fudge, 1969, Goldsmith, 1990) By contrast, Clements and Keyser, 1983, have no such internal structure to the syllable; it consists of simply C and V elements.

or coda with the preceding or following syllable respectively, in running speech) (see again, Clements and Keyser, 1983).

The approach followed here is to skirt the problems of syllabification by identifying the boundary between the syllable peak and whatever precedes it as the syllable boundary. For languages such as English, where there is both onset and coda, these are lumped together for the purposes of examining the question of accent alignment to syllable structure with respect to pitch prominence. The speech stream is then viewed as a quasi-periodic sequence of syllables, where the local period is the time from the start of one syllable peak to the next.

The question next to be addressed is how alignment of the salient points in the accent configurations discussed in this section to syllable structure affects the pitch prominence of those accents. Here, a common line is taken (see, e.g., Silverman and Pierrehumbert 1990), which is that it is the alignment of accent peaks with syllable peaks that is the important factor. The effect of the alignment of accent troughs is taken to be subsidiary to this factor and the time and frequency factors discussed in the previous subsection.

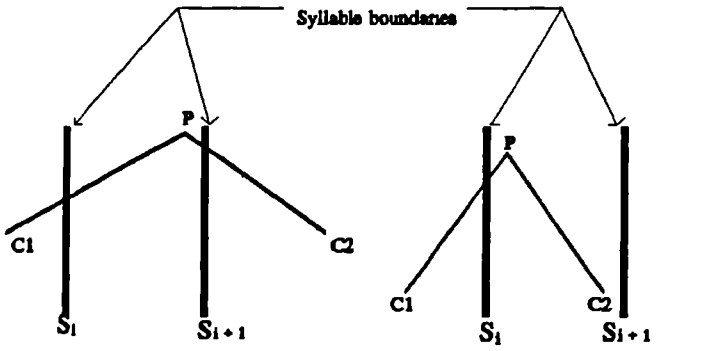
The comparison which is used as a basis for an initial estimate of the effect of peak alignment on pitch prominence is that between fall and rise-fall nuclear tones, using British English intonation as an example. In nuclear position, falling tones tend to have an accent peak fairly close to the syllable peak boundary (De Pijper 1983 modelled such falls by having the fall commence 50ms into the syllable peak, and the same observation is made by Silverman and Pierrehumbert 1990). Rising-falling tones, on the other hand, have accent peaks much later, at least at the end of the syllable peak[11] (De Pijper 1983, Pierrehumbert and Steele, 1987). Now, a qualitative estimate of the difference between the respective prominence of these two tonal types

---

[11] The peak in a rise-fall nuclear tone in English can appear as late as the end of the syllable peak of the post-nuclear syllable. In the current model, however, peaks which extend beyond the start of the syllable peak of the post-nuclear syllable have to be treated as part of a rising configuration, in which the trough preceding the peak becomes point P in the kind of schemata being used here, and the peak becomes point C2. This is an acceptable denotation of the points according to the general accent model in section 5.2.5.1.

can be made by informal observation and interpretation of the pragmatic gloss which is applied to them in auditory prosodic analysis. O'Connor and Arnold (1973), for instance, say that rise-falls (as in their 'Jackknife' pattern) impart a sense of impressedness and awe, and can have an intensifying function. From this and similar observations we can judge that rise-falls impart greater prominence, and that this is likely to be at least partly a function of greater pitch prominence, resulting from the change in the alignment of the accent peak.

Another observation that is relevant at this point is that head accents tend to have later peaks than nuclear accents (House, Johnson 1989, Silverman and Pierrehumbert 1990). This seems to run counter to what has just been said, since one would expect nuclear accents to be of greater prominence, ceteris paribus. In explanation, one possibility is that there is a compensating function independent of pitch which serves to give a net boost to the nuclear accent prominence. But a more likely reason for this conflict is the fact that the later-aligned peaks in head accents have, on average, more gradual onsets and offsets, whereas the early-aligned peaks in nuclear accents have, on average, steeper onsets and offsets; more specifically, the trough following a head accent will tend to be higher than that following a nuclear accent, ceteris paribus. This means that there is an interaction between the alignment of the accent peak and the values P/C1 and P/C2 as understood in the model developed so far.



Typical Head Accent Contour (peaked accent)

Typical Nuclear Accent Contour (peaked accent)

Leaving the question of this interaction aside for a moment, a first approximation to a quantitative estimate of the effect of peak alignment on pitch prominence would have a linear increase in contribution to pitch prominence as point P moved from the left boundary of one syllable peak to the left boundary of its successor.

However, since it is likely that there would be segmental material comprising syllable coda and/or onset in any arbitrary language before that left boundary of its successor, and since much of that material might be unvoiced, it is likely that some proportion of the contribution to prominence arising from the existence of the acoustically salient rise (in the case of the peaked accent) during the vowel of the syllable, concomitant with the delay of the peak, would be lost. A heuristic approach to tackling this problem, without incorporating any further information about syllable structure, is to have the rate of increase in the contribution to pitch prominence slow down as the peak approaches the left edge of the vowel of the following syllable. On average, this would give better estimates of the contribution to pitch prominence of peak alignment, though a more detailed model would take account of more precisely delineated local syllable structure[12].

A function suitable to account for this behaviour is the sine function. If only a quarter period of this function is used, with the distance of the peak from the left as a parameter, the following function is derived, whose value is a factor contribution to pitch prominence:

F6: $A = \sin(PI/2 * t_s(P)/(S_{i+1}-S_i))$
(where $t_s(X)$ is the position in time of point X (with minimum value 0 secs.) relative to an origin of the start of the syllable peak of the local syllable, S_i is that origin, and S_{i+1} is the position of the start of the syllable peak of the following syllable).

It was stated above that this factor contribution interacts with the ratios of the peak F0 value to trough F0 value. Examination of the behaviour of the

---

[12] For instance, in a CV-VC sequence of syllables, as in the English word "seeing", in nuclear position, the slowing down in the rate of increase in the contribution to prominence as the peak shifted leftward beyond a certain point might not be appropriate.

function indicates how this interaction should be formalised. If the accent peak is close to the beginning of the syllable peak, then the ratio P/C2 predominates in the contribution to pitch prominence, as the fall in the accent is most salient, and the ratio P/C1 has less contribution. At minimum, ($t_a(P)=0$), the peak-alignment factor contribution to pitch prominence as expressed in F6 can be considered to be minimal. As the accent peak gets further away from the start of the syllable peak, there is an almost (but decreasingly such a) linear increase in this factor contribution, until about three-quarters of the way through the distance between the syllable peak boundaries, when the rate of increase starts to level off. At this point, consonantal segments comprising the coda and then the onset of the next syllable would tend to appear (this is an intuitive statistical suggestion). To a first approximation, then, the factor contribution to pitch prominence expressed in F6 could be considered maximal beyond this three-quarter point. It is at and beyond this point that the ratio P/C1 would be expected to predominate in the contribution to pitch prominence, since the contour between the points P and C1 would then be more acoustically salient than that between points P and C2.

From these observations, the function A in F6 above could be thought to interact with the ratio P/C1, and its complement with P/C2. Now, it is clear that variation in P/C1 with the peak in the early position still results in variation in pitch prominence, and it is suggested that this variation is modulated by the time factor, regardless of the alignment of the peak, so the contribution of the peak alignment factor (which is zero when the peak is in early position, as far as the ratio P/C1 is concerned) must be made independent of the time factor contribution introduced in the last section. Yet in two of the models (1 and 2) detailed in that section, the frequency factor is not the simple ratio P/C, but a power function of it, so the ratio P/C cannot be used to modulate the independent time and peak alignment factors without some adjustment. Fortunately, this can be done straightforwardly, since

$$(P/C)^{2*0.167/(0.167+t(C)-t(P))} = P/C * (P/C)^{2*0.167/(0.167+t(C)-t(P))-1}.$$

Thus the frequency factors of Models 1 to 4 can be partitioned into two sub-factors: one interactive with the peak alignment factor, being the basic

frequency ratio P/C (or C/P), and one independent of it, being the frequency factor raised to the same power as the time factor (in Models 1 and 2) or unity (in Models 3 and 4).

Thus the appropriate form of the relevant functions is

F7: $A_{c1}(P) = \sin(PI/2 * t_n(P)/(S_{i+1}-S_i)$
F8: $A_{c2}(P) = 1 - A_{c1}(P)$

The functions A_{c1} and A_{c2} are peak alignment factor functions which jointly interact, respectively, with the basic frequency factor functions P/C1 (or C1/P) and P/C2 (or C2/P) (see the following).

The set of pitch prominence functions as derived so far for contour elements is as follows :

$F9_{m1}$: p(P,C1) = max(C1,P)/min(C1,P) * (F''(C1,P)*T(C1,P) + A_{c1})
$F9_{m2}$: p(P,C1) = max(C1,P)/min(C1,P) * (F''(C1,P)*T'(C1,P)+ A_{c1})
$F9_{m3}$: p(P,C1) = max(C1,P)/min(C1,P) * (T(C1,P) + A_{c1})
$F9_{m4}$: p(P,C1) = max(C1,P)/min(C1,P) * (T'(C1,P)+ A_{c1})

(where $(F''(C1,P) = (\max(C1,P)/\min(C1,P))^{2*0.167/(0.167+t(P)-t(C1))-1}$).

$F10_{m1}$: p(P,C2) = max(P,C2)/min(P,C2) * (F''(P,C2)*T(P,C2) + A_{c2})
$F10_{m2}$: p(P,C2) = max(P,C2)/min(P,C2) * (F''(P,C2)*T'(P,C2)+ A_{c2})
$F10_{m3}$: p(P,C2) = max(P,C2)/min(P,C2) * (T(P,C2) + A_{c2})
$F10_{m4}$: p(P,C2) = max(P,C2)/min(P,C2) * (T'(P,C2)+ A_{c2})

(where $F''(P,C2) = \max(P,C2)/\min(P,C2)^{2*0.167/(0.167+t(C2)-t(P))-1}$).

The set of pitch prominence functions as derived so far for accents is exemplified by the Model 1 function, as follows:

$F11_{m1}$: p(P,C1,C2) = max(C1,P)/min(C1,P) * (F''(C1,P)*T(C1,P) + A_{c1})
+ max(P,C2)/min(P,C2) * (F''(P,C2)*T(P,C2) + A_{c2})

(where F''(C1,P) and F''(P,C2) have the same interpretations as in F9 and F10).

Since the peak alignment functions $A_{P,C1}$ and $A_{P,C2}$ are complements of each other, summing to 1, and the value of the time factor function in each case is 1 in the canonical case (where the distance in time between P and C1 and between P and C2 is 0.167 seconds), it can be seen that there is a net increment in the value of the pitch prominence function as a result of the contribution of the alignment to syllable structure, compared with the function as expressed in function F5. For a peaked accent rising and falling over an octave over a period of 0.167 x 2 seconds, the value yielded by $F11_{M1}$ is 6ppu, whereas $F5_{M1}$ yields 4ppu. Given that the function in F5 would be as much appropriate for evaluation of the pitch prominence of an isolated accented vowel (such as one surprised form of the interjection "oh!", at least as spoken by an English person) as it is for a continuous AM tone, it might be though that these values should be the same. On the other hand, it might be thought that there is always a contribution to pitch prominence of peak alignment in speech, and that the absence of segmental material surrounding a vowel is compensated for in assessing pitch prominence. This question could be determined empirically, and provision is required for appropriate adjustment of the interactive alignment and time factors. This can be done by multiplying the sum of these factors by a constant expressing whether the effect of alignment to syllable structure enhances pitch prominence or not. For example:

$$F12_{M1}: p(P,C1,C2) = \max(C1,P)/\min(C1,P) * E * (F''(C1,P)*T(C1,P) + A_{C1}) + \max(P,C2)/\min(P,C2) * E * (F''(P,C2)*T(P,C2) + A_{C2})$$

(where E is the enhancement constant).

If E has a value of 1, F12 is clearly equivalent to F11. To make the value of the function equivalent to that of F5, E should have a value of 0.667. It is conjectured here that the appropriate evaluation of the pitch prominence function has F5 and F12 equivalent (that is, there is no overall increment in the function provided by the mere fact of alignment to syllable structure), and so the value E=0.667 is used in later computations.

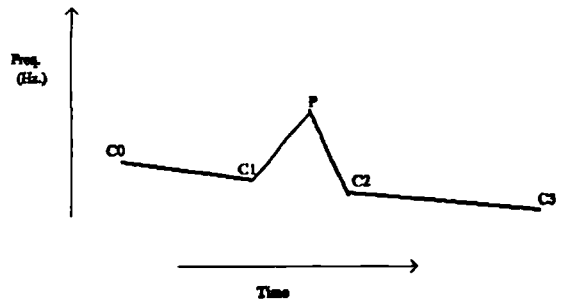
5.2.8 Prominence as a function of unaccented context

So far in this chapter, consideration has only been given to the pitch prominence of accent configurations. In this subsection, the contribution is considered of the configurations which motivated the current analysis, viz.

those of F0 contours before and after accent configurations, comprising unaccented material.

Unaccented material is characterised by having shorter syllables, in which segmental coarticulation effects are proportionately larger. This can be seen in any of the natural F0 contours of Chapter 2, in which the F0 movements on unaccented syllables can be as steep as those on accented syllables, though less extensive (and associated with lower speech amplitude in the general case). In this initial approximation to a model of the pitch prominence of individual accents, the context which is supplied by unaccented material has to be simplified in some way. Therefore, the following discussion will assume that a straight-line stylisation of the unaccented material has been performed, perhaps by fitting a regression line. This does run counter to the principle of trying to determine pitch prominence as a function only of F0 variation physically present in the F0 contour, but is justified in being a compromise in a step on the way to determining such a function. Thus, the first level of abstraction referred to in footnote 2 of Chapter 2 (type (b)) is taken as given in this analysis.

Taking the peaked accent as a reference point, a schema for the accent which includes unaccented syllables is as follows:



It is contour elements such as the stretches from C0 to C1 and from C2 to C3 which the Local Declination Hypothesis considers to be the vectors of the

declination effect. That is, if those contour elements are declining, then the declination effect is likely to be observed on any immediately subsequent accented syllable; to elicit equal prominence, the peak of the latter accent will have to have a lower F0 value than point P in the figure.

It will be seen in what follows that the essence of the Local Declination Hypothesis can be maintained - that the declination effect is to a large part attributable to local variation in the unaccented material - but that as it is stated, it is too simple a picture to paint.

First, though, an estimate of the contribution to pitch prominence of individual unaccented contour elements has to be made. There is no reason to suppose that the basic function determining pitch prominence for these should be different from that determining the pitch prominence of constituent contour elements of accents. However, since no cognisance is necessarily made of syllable structure in determining the salient points of the contour element, and since unaccented contour elements span more than two syllables in many cases, the peak alignment factor and the associated enhancement constant should be removed from the function. This makes the set of pitch prominence functions for an unaccented contour element as follows:

$F13_{M1}$: p(C,C') = max(C,C')/min(C,C') * F''(C,C') * T(C,C')
$F13_{M2}$: p(C,C') = max(C,C')/min(C,C') * F''(C,C') * T'(C,C')
$F13_{M3}$: p(C,C') = max(C,C')/min(C,C') * T(C,C')
$F13_{M4}$: p(C,C') = max(C,C')/min(C,C') * T'(C,C')
(where C is the start point and C' the end point of the unaccented contour element[13]).
This set of functions is adjusted in circumstances described below in section 5.3 .

The question now arises how the contour elements comprising the unaccented context (henceforth, the "context") combines with those comprising the accent to produce a composite pitch prominence function for the whole configuration (such as C0–C3 in the diagram above). It seems that the

[13] Note that for a prepeak contour element, the start point is later in time than the end point. This is because the constituent points are viewed backwards from the peak.

simplest combination, addition of the constituent prominence values, is appropriate.

This can be seen from the following observation. If two sequences of "ar" or "ah" (open back unrounded) vowels are synthesised with peaked F0 contours comprising equal frequency and time excursions, the first sequence being based on increasing duration of the contour element C0-C1 from 0 seconds to 0.25 seconds (contour element C2-C3 being fixed at 100ms) and the second sequence being based on increasing duration of the contour element C2-C3 from 0 seconds to 0.25 seconds (contour element C0-C1 being fixed at 100ms), the following impression is given. The prominence of the vowels in the first sequence increases up to a point at which C0-C1 is of about 120ms duration, beyond which it levels off and then is reduced at a duration a little greater than 170ms. The prominence of the vowels in the second sequence increases up to a point at which C2-C3 is of about 100ms duration, beyond which it levels off and then gradually reduces. These are subjective observations made during an informal experiment conducted by the author, to provide some confirmation of clear indications made during research into synthesis-by-rule of intonation contours (Johnson 1990[14]) that pitch excursions without any flattened tail don't sound as extensive as those with.

Thus, the composite pitch prominence function (using Model 1 as an example) for accent and context is as follows :

$$F14_{M1}: p(P,C0,C1,C2,C3) = \max(C1,P)/\min(C1,P)*E*(F''(C1,P)*T(C1,P) + A_{C1})$$
$$+ \max(P,C2)/\min(P,C2)*E*(F''(P,C2)*T(P,C2) + A_{C2})$$
$$+ \max(C1,C0)/\min(C1,C0) * F''(C1,C0) * T(C1,C0)$$
$$+ \max(C2,C3)/\min(C2,C3) * F''(C2,C3) * T(C2,C3)$$

The function is not quite complete. Just as the pitch prominence of a contour element is constrained by its duration, so is it constrained by the amplitude of the associated speech. The speech has at least to be of sufficient amplitude

[14] It should be emphasized that the model of intonation developed here is different from that discussed in the cited report. It does, however, continue investigation of one feature of that model, which is that declination is a local feature of individual accent configurations.

for pitch to be detectable. Here, this factor is not accounted for in detail, it being considered that a separate amplitude prominence function, interactive with the pitch prominence function, is required for a proper account. An adjustment is made, however, by introducing a relative amplitude factor, which is a normalised variable specifying the relative mean RMS amplitude values[15] for the two parts of the context, and the accent taken as a whole. The value of this variable for one of those three constituents is therefore 1 (the constituent with maximum relative mean RMS amplitude), and the others have values less than or equal to it[16]. Because it is a measure of relative amplitude, it need not be specified in the model for the accent constituent.

Thus, the complete pitch prominence function (using Model 1 as an example) is as follows :

$F15_{M1}$: p(P,C0,C1,C2,C3) = max(C1,P)/min(C1,P)*E*(F''(C1,P)*T(C1,P) + A_{C1})
+ max(P,C2)/min(P,C2)*E*(F''(P,C2)*T(P,C2) + A_{C2})
+ max(C1,C0)/min(C1,C0) * F''(C1,C0) * T(C1,C0) * $R_{C1,C0}$
+ max(C2,C3)/min(C2,C3) * F''(C2,C3) * T(C2,C3) * $R_{C2,C3}$

(where $R_{C,C'}$ is the relative amplitude factor for the contextual contour element C-C').

This can be expressed in words as follows:

Pitch prominence of accent (P,C1,C2) and context (C1,C0),(C2,C3) configuration = pitch prominence of prepeak accent contour element
+ pitch prominence of postpeak accent contour element
+ pitch prominence of prepeak context contour element
+ pitch prominence of postpeak context contour element

---

[15] The RMS amplitude being computed over non-overlapping 40ms time windows.

[16] The normalisation is not done so that the lowest amplitude constituent has a relative amplitude value of zero, but so that it has a value in the same proportion to the maximum value as when unnormalised.

5.3 PREDICTING PEAK HEIGHT

5.3.1 Introduction

Given the measure of individual accent pitch prominence proposed in the previous section, it is possible to construct a function which returns an F0 value for the peak of a non-initial accent in a tone-unit, using the pitch prominence of contour elements local to the preceding accent and surrounding context as parameters, which has the same pitch prominence as the preceding accent. The peak F0 value of a following accent of a desired degree of pitch prominence can thus be predicted on the basis of an adjustment to this equal-prominence value.

5.3.2 A model for predicting peak height

Starting with the relationship between the two accents

F16a: $p(P_2,C0_2,C1_2,C2_2,C3_2) = J * p(P_1,C0_1,C1_1,C2_1,C3_1)$

that is

F16b: $\max(C1_2,P_2)/\min(C1_2,P_2)*E*(F(C1_2,P_2)*T(C1_2,P_2) + A_{c1})$
$+\max(P_2,C2_2)/\min(P_2,C2_2)*E*(F(P_2,C2_2)*T(P_2,C2_2)+A_{c2})$
$+ \max(C1_2,C0_2)/\min(C1_2,C0_2)*F(C1_2,C0_2)*T(C1_2,C0_2) * R_{c1,c0}$
$+ \max(C2_2,C3_2)/\min(C2_2,C3_2)*F(C2_2,C3_2)*T(C2_2,C3_2) * R_{c2,c3}$
$= J * (\max(C1_1,P_1)/\min(C1_1,P_1)*E*(F(C1_1,P_1)*T(C1_1,P_1) + A_{c1})$
$+\max(P_1,C2_1)/\min(P_1,C2_1)*E*(F(P_1,C2_1)*T(P_1,C2_1)+A_{c2})$
$+ \max(C1_1,C0_1)/\min(C1_1,C0_1)*F(C1_1,C0_1)*T(C1_1,C0_1) * R_{c1,c0}$
$+ \max(C2_1,C3_1)/\min(C2_1,C3_1)*F(C2_1,C3_1)*T(C2_1,C3_1) * R_{c2,c3}$
)

(where J is the factor difference in prominence between the two accents, and is 1 for equal prominence)

it is possible to derive, assuming that $\max(C1_2,P_2) = P_2$, and $\max(C2_2,P_2) = P_2$,

F17: $P_2 = C1_2 * C2_2 \,/\, (E * (C2_2 * (F(C1,P) * T(C1_2,P_2) + A_{c1})) + C1_2 * (F(C2,P) * T(C2_2,P_2) + A_{c2})))$

$* (J * (\max(C1_1,P_1)/\min(C1_1,P_1) * E * (F(C1_1,P_1) * T(C1_1,P_1) + A_{c1}) + \max(P_1,C2_1)/\min(P_1,C2_1) * E * (F(P_1,C2_1) * T(P_1,C2_1) + A_{c2}) + \max(C1_1,C0_1)/\min(C1_1,C0_1) * F(C1_1,C0_1) * T(C1_1,C0_1) * R_{c1,c0} + \max(C2_1,C3_1)/\min(C2_1,C3_1) * F(C2_1,C3_1) * T(C2_1,C3_1) * R_{c2,c3}) - \max(C1_2,C0_2)/\min(C1_2,C0_2) * F(C1_2,C0_2) * T(C1_2,C0_2) * R_{c1,c0} - \max(C2_2,C3_2)/\min(C2_2,C3_2) * F(C2_2,C3_2) * T(C2_2,C3_2) * R_{c2,c3})$

(where F(C1,P) and F(C2,P) are identical to $F(C1_1,P)$ and $F(C2_1,P)$ respectively. It should be noted that the use of frequency factor functions for the first accent prominence function in place of those for the second accent prominence function is necessary in order to be able to derive P_2 without recourse to numerical analytic methods in the case of Models 1 and 2 (see section 5.2.7); in the case of Models 3 and 4, these functions just have a value of unity. However, it is justified only in those cases when the shape of the accent in accent configuration 2 is identical to that in accent configuration 1 (as in Figure 5.9). In the general case, iterative numerical analytic methods must be used to determine the value of P_2 for Models 1 and 2. For Models 3 and 4, F17 can be used with F(C1,P) and F(C2,P) set to 1).

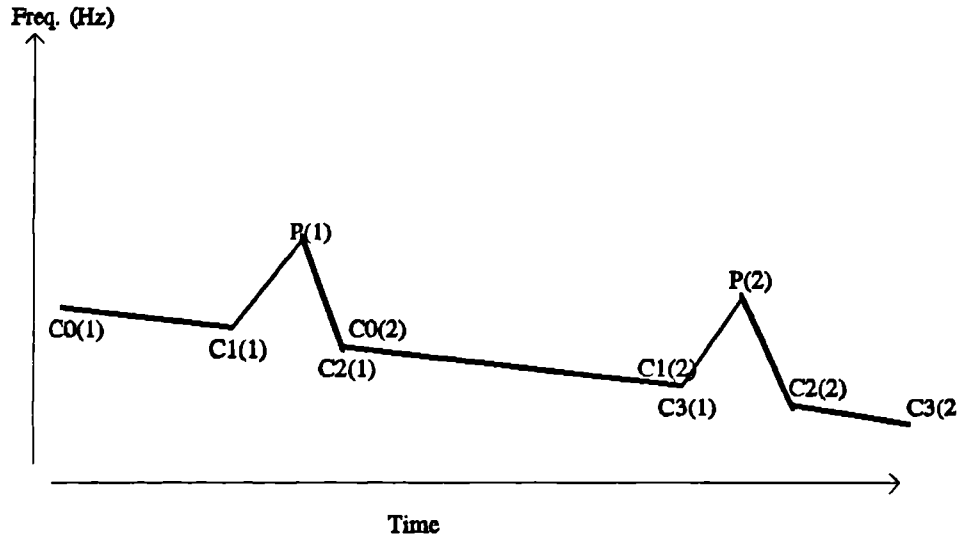


Figure 5.9 A contour with two peaked accents. Salient points used in the pitch prominence model are indexed according to the accent configuration to which they belong.

Figure 5.9 depicts the sort of contour used by Terken in his experiments reported on in Chapter 4. One of the things that has to be decided on in identifying the salient points in the contour and their constituency is whether there should be any overlap in the context, and how extensive it should be. Here, contour element C2-C3 of the first accent's postpeak context overlaps completely with contour element C0-C1 of the second accent's prepeak context - this can be considered a manifestation of tonal 'sandhi', although it is not a typical context for its occurrence.

5.3.3 Overlap of contextual contour elements in adjacent accent configurations

Now, according to F17, the prominence factor contributed by $C2_1$-$C3_1$ and that contributed by $C0_2$-$C1_2$ more or less cancel each other out in the computation of P_2 (as the relative amplitude values may not be quite the same, the cancellation may not be complete). This means that the local interaccentual context would have virtually no influence in determining the peak height of the second accent. For that not to be the case, either there should be considered to be no overlap between contextual contour elements when there is the possibility for it, or there should be found cause for

adjustment of the prominence functions of contextual contour elements, either in general, or in the case of the those that overlap.

It seems unrealistic to prohibit overlap of contour elements when there is the possibility of it. It is clear that the speech stream is partly the product of a sequence of overlapping supraglottal articulatory gestures, and must be decoded as such. Equally clearly, the contribution of excitation to the speech stream is the product of a sequence of overlapping laryngeal gestures. Although the relationship between such gestures and the contour elements identified in the model developed here is not precise, there is some support for the suggestion that contour elements should overlap for a period of at least the order of the interaccentual stretch in Terken's basic stimulus (chapter 4) – at least different laryngeal gestures can overlap for that period of time (see Fig 2.17). Thus, in the absence of direct evidence against it, the overlapping of at least contextual contour elements is allowed.

If overlapping elements are allowed, some adjustment has to be made in the pitch prominence functions for any elements which do overlap. This can be seen from the following: if the duration of a contextual contour element is short (of the order of 200ms or less), and the F0 ratio in its basic frequency factor is sufficiently large, it begins to have parameter values appropriate to an accentual contour element. This is one of the essential features of the current model; the basic model for accent and context is the same, reflecting the fact that accentuation can develop from any part of the context. Now if overlap took place between adjacent accent contour elements (this could theoretically occur in English intonation, say, when the rising contour element of a fall-rise tone merges with the prepeak contour element of a following step accent), the current model would predict that the prominence of that rising contour element was effectively zeroized by the overlap. Since it is taken as axiomatic that accentuation boosts prominence, this is an unacceptable state of affairs. Consequently, the model would have to be adjusted; an appropriate way of doing this is to let the basic frequency ratio for overlapping contour elements be simply that between the first and second F0 values supplied to the pitch prominence function, rather than that between the maximum and minimum of these values. In the case of the rise part of the fall-rise overlapping with the prepeak contour of a step accent, this would mean that the frequency factor for the rise part would be less than

1 and that for the prepeak contour element greater than 1 (as the first F0 value supplied to the pitch prominence function is the peak value and the second the trough value, for a stepped up to step accent), so that the latter would predominate in terms of pitch prominence.

5.3.4 A revised model for predicting peak height

In general, the probability of overlap of adjacent accentual contour elements is fairly low. The illustration in the previous paragraph serves only to underline that if overlap is allowed, the basic frequency factors (the ratio between maximum and minimum of the first and second supplied F0 values) of the overlapping contour elements should be adjusted to become the ratio simply between the first and second supplied F0 values. The probability of overlap of adjacent contextual contour elements is, on the other hand, quite high. Thus it is these elements that are more amenable to the adjustment to the basic frequency factor suggested. Now the model in F17 can be generalised to allow for the prediction of peak height of the successor to any arbitrary accent in a sequence. The most succinct way of representing the likelihood of adjustment to the contextual contour element pitch prominence factors as a result of overlap is to specify that their basic frequency factor is so adjusted without condition. Thus:

$$
\begin{aligned}
\text{F18: } P_{i+1} = {} & C1_{i+1} * C2_{i+1} \,/\, (E * (C2_{i+1} * (F(C1,P) * T(C1_{i+1}, P_{i+1}) + A_{c1})) \\
& \qquad + C1_{i+1} * (F(C2,P) * T(C2_{i+1}, P_{i+1}) + A_{c2}))) \\
& * (J * (\max(C1_i, P_i) / \min(C1_i, P_i) * E * (F(C1_i, P_i) * T(C1_i, P_i) + A_{c1}) \\
& \qquad + \max(P_i, C2_i) / \min(P_i, C2_i) * E * (F(P_i, C2_i) * T(P_i, C2_i) + A_{c2}) \\
& \qquad + C1_i / C0_i * F(C1_i, C0_i) * T(C1_i, C0_i) * R_{c1,c0} \\
& \qquad + C2_i / C3_i) * F(C2_i, C3_i) * T(C2_i, C3_i) * R_{c2,c3} \\
& \qquad) \\
& \qquad - C1_{i+1} / C0_{i+1} * F(C1_{i+1}, C0_{i+1}) * T(C1_{i+1}, C0_{i+1}) * R_{c1,c0} \\
& \qquad - C2_{i+1} / C3_{i+1} * F(C2_{i+1}, C3_{i+1}) * T(C2_{i+1}, C3_{i+1}) * R_{c2,c3} \\
&) \\
& \text{(for i=1 to n-1, where n=number of accents in the tone unit).}
\end{aligned}
$$

A more detailed and accurate model would make the use of the ratio in the terms for the context conditional on the existence of overlap, and might adjust those terms according to the degree of overlap.

If the ratio in any of the terms for the context is less than 1, the shape of the functions determining the variation of pitch prominence factor with time for Models 1 and 2 (see Figures 5.1-6) changes, because of the use of a base in the frequency factor which is less than unity. The result is that the frequency factor becomes asymptotically increasing rather than asymptotically decreasing, and the composite factor consequently peaks at a lower value in the case of Model 1, and effectively peaks only very late in the case of Model 2. This behaviour can be seen in Figures 5.10 and 5.11 .

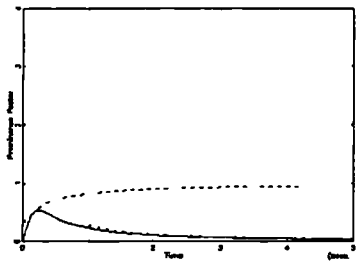


Figure 5.10 Pitch prominence factor variation with time - Model 3, with base < 1.

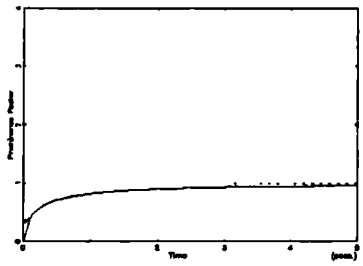


Figure 5.11 Pitch prominence factor variation with time, with base less than 1 - Model 4.

5.3.5 A unitary model of the production and perception of intonation

The model in F18 has been developed on the basis of a model of prominence for individual accents which uses perceptual criteria for the scaling of their pitch prominence, but it forms the basis of a model of speech production. In this sense, it is a model of the perceptual determination of the production of intonation, and predicts, generally, that there will be auditory feedback involved in that process.

To exemplify the operation of the model, the case of a peaked accent sequence is considered. The model can be interpreted as stating that when the F0 value of an impending accent is computed in the process of producing an F0 contour, the following are used as parameters: the prominence value of the preceding accent, expressed in terms of its peak F0 and the F0 values of the troughs surrounding it, the prominence value of its preceding (prepeak) context, expressed in terms of the start and end F0 values of that

context, the prominence value of its following (postpeak) context, expressed also in terms of its start and end F0 values, the prominence value of the prepeak context of the impending accent (which could be coterminous with that of the postpeak context of the preceding accent), and the prominence value of the postpeak context of the impending accent. The computation of the F0 value is fairly straightforward; the prominence parameters are combined and modulated by, reorganised, that part of the putative prominence function of the impending accent which is not its peak F0 value.

The computation can be considered from the point of view of the speaker from any point between the end of the accent just produced and the start of the impending accent. In these positions, there are two different classes of variable, those for which values can be determined by feedback, and those which are predictions of variables whose actual value in production is yet to be determined. The former can be called 'lookback' variables, and the latter 'lookahead' variables. The lookahead variables consist of the trough F0 value just prior to the peak of the impending accent, and the trough F0 values just after that impending peak. This situation is summarised in Figure 5.12 .

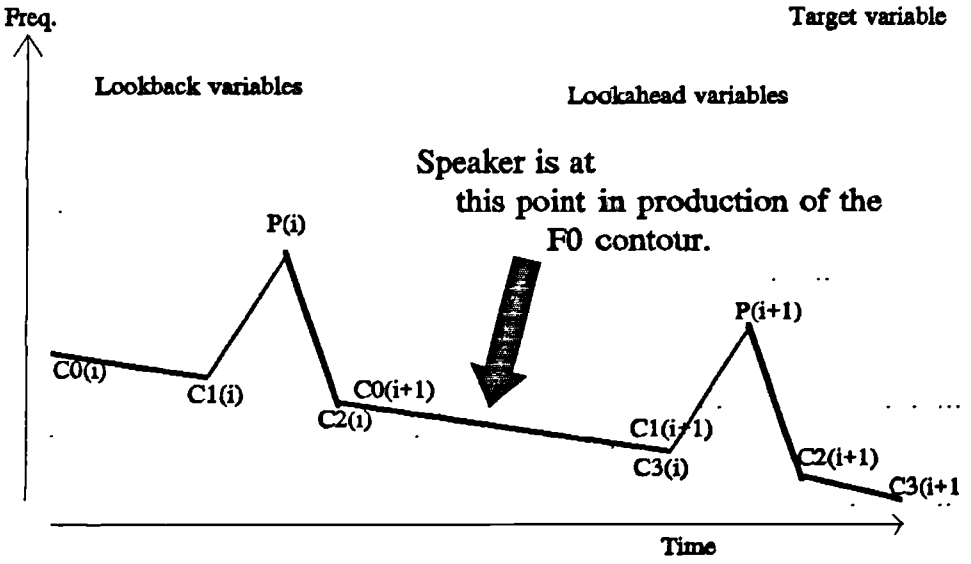


Figure 5.12 Different classes of variable in the process of producing an F0 contour with perceptual constraints.

The peak F0 of the impending accent is presented as the target variable in this figure. Now, there is a sense in which all of the variables beyond the arrowed point are target variables. If the F0 value of the lookahead variable C1(i+1) (= $C1_{i+1}$) is not actually achieved, then adjustments would have to be made in the values of the other lookahead variables for the target F0 value to impart the desired degree of prominence. If the F0 values of the lookahead variables C2(i+1) and C3(i+1) are not achieved, then it is of course by that time too late to make any adjustments to maintain the desired prominence on the peak, so that the actual prominence value of the resulting accent configuration may be different from that intended. At the same time, any alterations to the lookahead variables as a result of their having actually been produced or predictions of their value having changed, because of a change of strategy or for any other reason, could end up in an adjustment of the peak F0 value on P(i+1). In this sense, there is scope for the lookahead variables interacting in a network of models of the form of F18, each of them taking the place of $P1_{i+1}$ as the dependent variable in the model, their final values being determined by mutual adjustment.

If the single model is considered by itself, with the peak of an upcoming accent as a target variable, the variables of the context, in particular the interaccentual context, can be used as elements in a local feedback model, in which the slope of a contextual contour element is the controlled variable. Under this interpretation, a typical strategy in that part of the speaker's control process would be to have as a primary target the trough prior to an upcoming peak (point $C1_{i+1}$ in F18[17]), which, with the other variables in the model being either predicted, or known by feedback, would determine a particular target F0 value of the upcoming peak for a particular prominence relationship with its predecessor. An initial slope from point $C0_{i+1}$ to point $C1_{i+1}$ would then be set up as the reference slope in a feedback model, given the predicted duration between the two points[18], and the error between the

---

[17] This could be computed in the same way as P_{i+1} in F18, after appropriate algebraic manipulation of the model.

[18] It is likely that the value of the reference slope would also be adjusted, particularly in spontaneous speech, on the basis of an increase or decrease in the actual amount of speech material between the accents, compared with

actually occurring slope and the reference slope would be used to adjust those productive processes which determine the slope - notably the joint balance between contraction and relaxation of the expiratory and inspiratory respiratory and extrinsic and intrinsic laryngeal musculature. In this typical case, the reference slope is negative; that is, the target F0 contour between peaks is declining, and in this sense the model predicts the possibility of the auditory control of declination.

The model in F18 is still valid, however, in the presence of a concomitant model of the productive physiological determination of the production of intonation. That is, a speaker may exert varying degrees of control over the variables in the model of F18. If the passive elements in the physiological mechanisms involved in the production of speech are allowed to predominate, then one of the effects may be that subglottal pressure decline has a strong influence on the course of F0, notably during unaccented stretches when glottal resistance tends to be lower because of the reduced tension resulting from reduced Cricothyroid, Vocalis and Lateral Cricoarytenoid muscle activity. Yet the speaker can still have retrieved the values of the variables through auditory feedback mechanisms (or through lookahead) and determine the peak target value of an imminent accent that reflects a desired prominence relationship with a preceding accent, for which appropriate adjustments in the productive mechanisms can be made if necessary. Those passive elements could predominate even to the extent of the peak height of an accent or number of accents in the intonation contour being largely determined by them, yet a speaker could still make adjustments for the height of a subsequent accent to reflect a desired prominence relationship with its predecessor on the basis of values in the model of F18 being monitored during the course of speech.

As for the perception of intonation, the model predicts that a scaled estimate of the prominence of an individual accent can be made on the basis of its configuration (accent proper plus context) and the ratios between the end points of its constituent contour elements. An estimate of relative prominence between two adjacent accents is thus derivable simply as the ratio between their respective prominence values.

---

the original prediction.

5.3.6 Verification of the model

In order for it to be of any use, the model has to work, or show the potential for working, for the intonation of a particular language. Since it was developed by iterative methods over many stages using English intonation as an intuitive reference point, it carries the danger of having built-in language-specific constraints. It is demonstrated here, however, using reiterant speech with intonation patterns applicable to both English and Dutch, and the model is sufficiently general to be applicable after possible modification to other intonation and pitch accent languages.

5.3.6.1 Terken's experimental data

A first test of the model can be made against the values for P2 in Table 4.1 of Chapter 4, that is the F0 values for the peak of the second peaked accent judged to be of equal prominence to the first in the second of the second pair of Terken's experiments reported in that chapter. Of the four models of prominence factor variation with time depicted in Figures 5.1-6, Model 4 gives the best match to these values. Table 5.1 lists these values and the values for P1 for which that model used in F18 (with J=1) predicts equal prominence.

All of the other three models predicted a narrower range of P2 values. Figure 5.13 shows the contour computed with the predicted P2 value for the highest contour of the eleven in Terken's experiment, using Model 1 in F18. Figure 5.14 shows the contour with the predicted P2 value for that same highest contour using Model 4 in F18.

Table 5.1 F0 values for an equal prominence second peak (see contour in Figure 5.9) predicted by the model in F18, using Model 4 as the basic prominence factor variation model. Comparison columns from Table 4.1.

Observed P1	Predicted P2 (equal prominence)	Observed P2 (equal prominence)
156	131	136
152	130	132
145	125	129
141	126	127
135	123	122
130	120	121
125	116	117
120	114	114
116	112	111
112	110	107
108	108	104

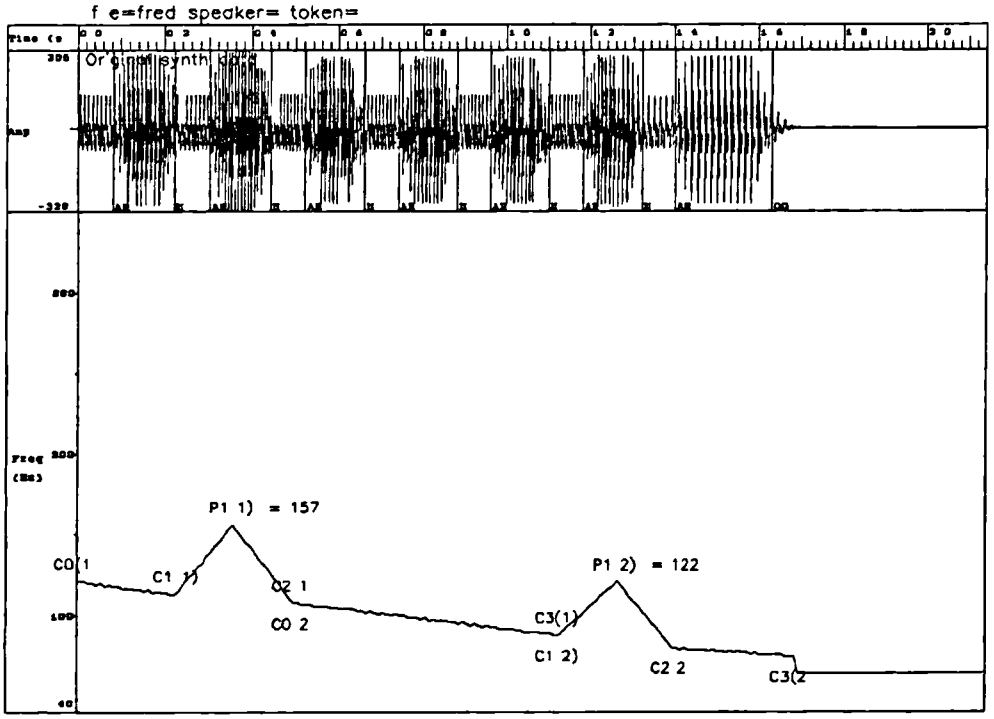


Figure 5.13 - F0 contour conforming to F18 (model 1, J=1) and associated synthetic speech pressure waveform for two peaked-accent rendition of the utterance "ma MA ma ma ma MA ma".

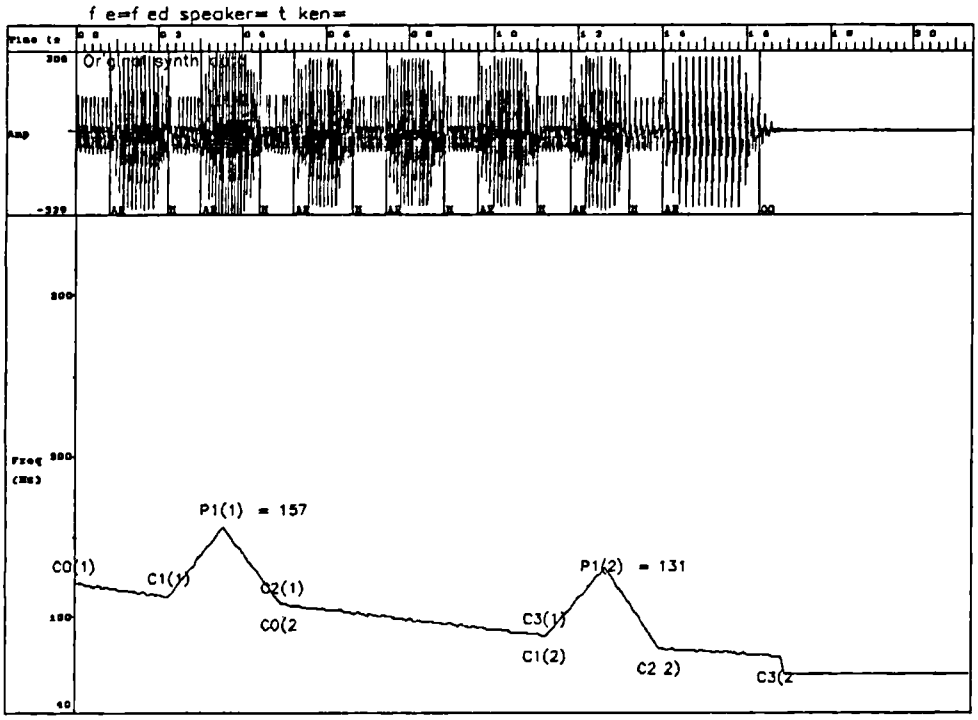


Figure 5.14 As Figure 5.13, but using Model 4 in F18.

It is clear that the relative amplitude factor has a minimal effect in these contours, since the syllables are all of approximately equal amplitude.

5.3.6.2 Predicted peak height following a falling head

Given that the model is sensitive to the slope of unaccented material, it is of interest to observe what peak height it predicts when the unaccented material doesn't follow an apparent baseline, as it does in Figures 5.13 and 5.14. Of particular interest is the configuration in which a falling head follows peak 1 and precedes peak 2.

According to the Global Declination Hypothesis, the prominence of peaks 1 and 2 should be computed by reference to a global declining baseline extending over the whole tone-unit; therefore, the F0 value on a second peak of equal prominence to the first peak should be the same, regardless of the slope of the intervening unaccented material (particularly when there are additional cues to the existence of such a global baseline in the start and end points of the whole contour) and will be less than the F0 value on the first peak.

The Local Declination Hypothesis, on the other hand, predicts that it is the physical (or audible) existence of declination between the two accents (and effectively, the existence of final lowering) which determines the F0 value on a second peak of equal prominence to the first peak, and that if there is declination (and/or final lowering) present, the second peak will be lower than the first. An elaboration of this hypothesis is that the more declination there is between the peaks, the lower the second peak will have to be for an equal prominence judgment to be made of the two peaks.

In figure 5.15 is a contour with a falling head, whose second peak has been computed according to F18 (using Model 3 as the basic model of pitch prominence factor variation). The predicted peak height is lower by 4Hz than that generated for Model 3 for the sorts of contour appearing in Figures 5.13 and 5.14 (and is thus tending towards the correct value for this configuration in English, according to the intuitions of the author). In Figure 5.16 is the contour generated by Model 4, which shows a contrary tendency, as it has a higher peak than that produced for the contour in Figure 5.14.

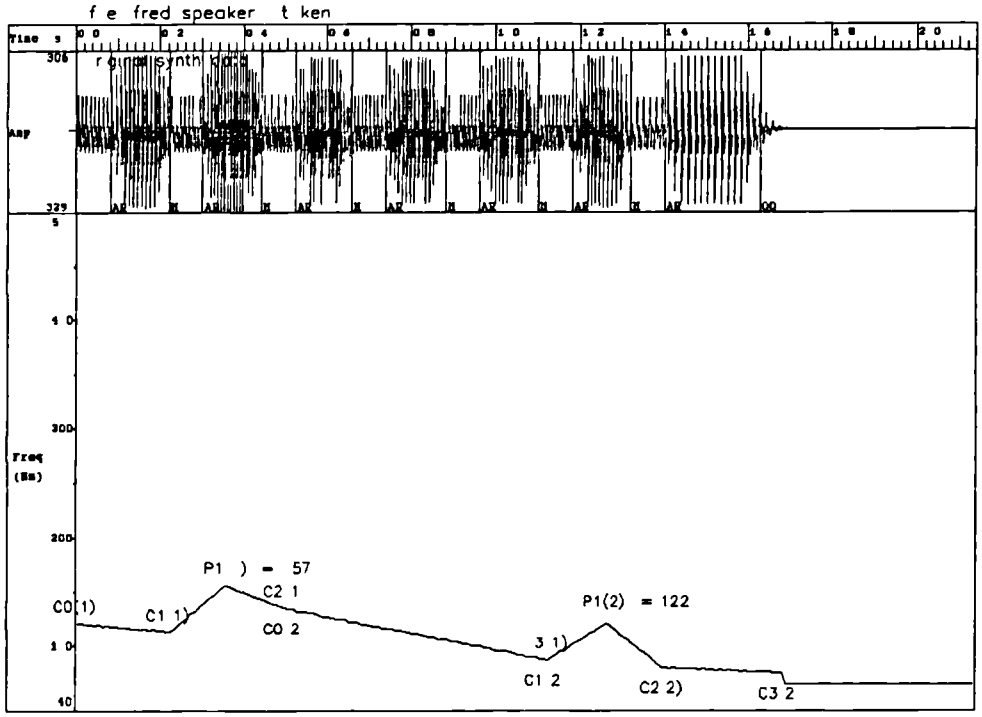


Figure 5.15 F0 contour with falling head, 2nd. peak computed by F18 (Model 3).

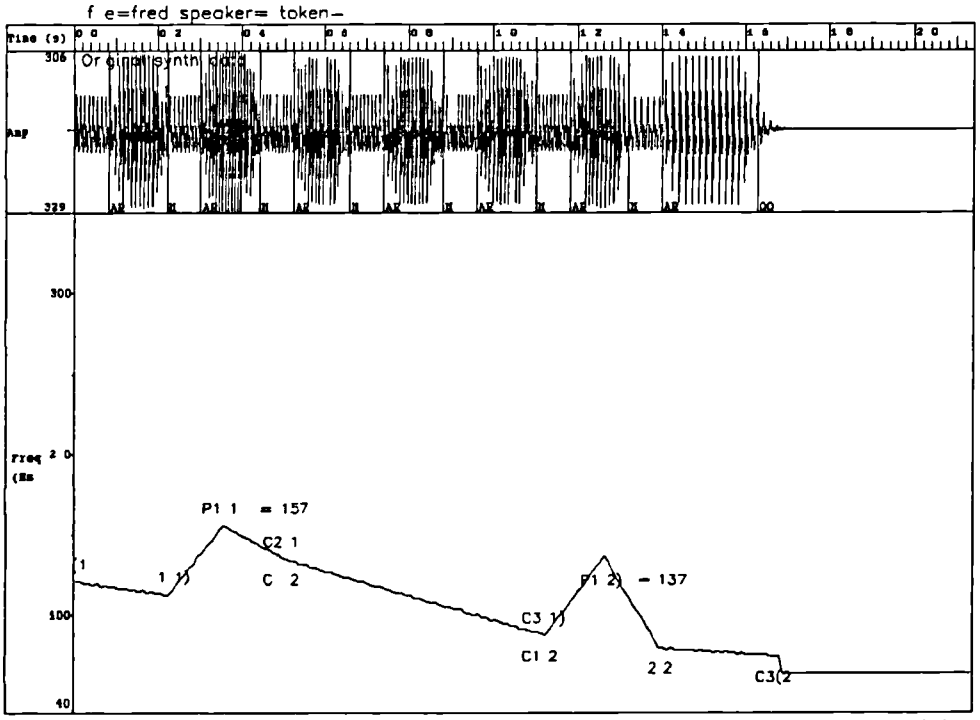


Figure 5.16 F0 contour with falling head, 2nd peak computed by F18 (Model 4).

It could be objected here that point $C0_2$ in Figures 5.15 and 5.16 is too theoretical a construct, and is insufficiently salient to warrant marking as the boundary of a contour element. It can be accepted that this is possible (although the observations of section 5.2.4 on the gating of F0 contour information at a rate of c. 6Hz are relevant here). As a consequence, the behaviour of the model when points $C3_1$ and $C0_2$ are superimposed on point $C1_2$ should be observed.

Figures 5.17 and 5.18 show the contours with equal-prominence second peaks computed according to Models 3 and 4 respectively, in which that immediately post-peak salient point is removed, and the first accent is effectively modelled as having a steep prepeak contour element and a more gradually declining longer postpeak contour element. It can be seen that the predicted peak value is lower in each case than when an identifed interaccentual contour element is interspersed between the accents, but that relative to the case when the effective declination between the accent peaks is shallower, the behaviour is the same; that is, the steeper the observable declination between the accents, the greater is the prominence-F0 ratio on the second accent. The effect is very much stronger for Model 3 than for Model 4 (probably too strong). The difference in second peak height for Model 4 between the case with shallow interaccentual declination and the case with steep interaccentual declination according to this specification is only 2Hz. This makes model 4 more consistent with the Global Declination Hypothesis.

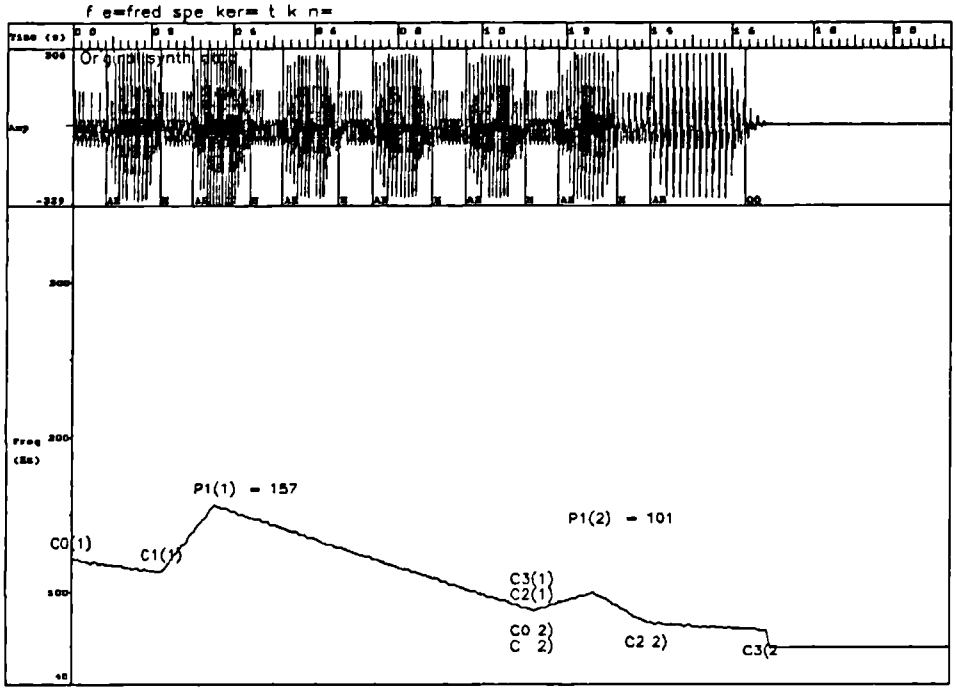


Figure 5.17 F0 contour with falling head modelled by accentu al contour element with gradual slope. Equal prominence F0 value on 2nd. peak computed according to F18 (Model 3).

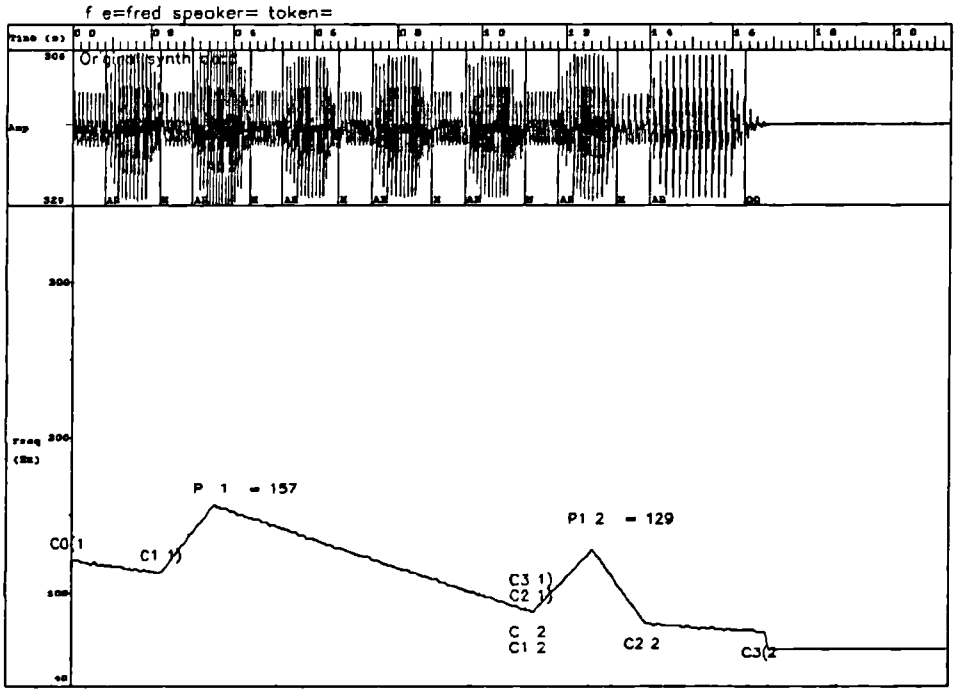


Figure 5.18 F0 contour with falling head modelled by accentu al contour element with gradual slope. Equal prominence F0 value on 2nd. peak computed according to F18 (Model 4).

5.3.6.3 Predicted peak height following a rising head

Since the F0 value of an equal-prominence peak is predicted to be unaffected by variation in slope of the interaccentual contour by the Global Declination Hypothesis, a change in its sign, too, should not alter the peak F0 value. On the other hand, the Local Declination Hypothesis actually makes no prediction as to the F0 value of an equal-prominence peak when the F0 contour between the accents is rising. The corollary to the hypothesis in Chapter 4 states that if there is no local declination between the two accent peaks, then they are of equal prominence, all other things being equal. The final rider means, inter alia, that only if the shapes of the accent peaks are the same, as they would be if the accents were surrounded by level pitch contours, would the equal-prominence peaks have the same F0 value on them. Now, this is not the case if the interaccentual stretch is either rising or falling, since one of the two accents has a reduced height pre- or post-peak contour element in it in those circumstances. Therefore, the hypothesis could be seen to predict a different prominence-F0 ratio for peaks on accents following different contextual contour elements, as a function not specifically of the physical presence of declination or inclination in those contextual contour elements, but of the resultant variation in accent shape[19].

However, this is against the general thrust of the Local Declination Hypothesis, as it argues against the primacy of Part 1 of it, that declination has to be physically manifested in some part of the intonation contour for the prominence-F0 ratio to be adjusted on adjacent accent peaks. Consequently, it could be considered appropriate to amend the various parts of the hypothesis, so that local inclination as well as declination is predicted to have an effect on the prominence-F0 ratio of the accent peaks:

LDH (Local Declination Hypothesis - a revised version)

PART 1

Declination or inclination has to be physically manifested in some part of the intonation contour to have an effect on the prominence ratins elicited for the following accent peaks.

---

[19] For this reason, the unamended version of the general model, expressed in F17, could be seen to be consistent with the Local Declination Hypothesis (depending on which of Models 1-4 was used within it). But as discussed in the following, the spirit of the hypothesis is contradicted by that position.

PART 2

There is a local perceptual declination effect, such that if a local declination is present between two accents, it increases the prominence-F0 ratio on the second accent. There is also a local perceptual inclination effect, such that if a local inclination is present between two accents, it reduces the prominence-F0 ratio on the second accent.

COROLLARY

If there is no declination, and no inclination, between two equally high accent peaks, the two accent peaks are of equal prominence, ceteris paribus.

The veracity of this hypothesis is clearly something that can be determined empirically. It is considered here that even if the general principle of the Local Declination Hypothesis is correct, it could still turn out that the effect on the prominence-F0 ratio of surrounding accents of local variation in the slope of contextual contour elements is still restricted to the case of declining slope, this effect being mediated, perhaps, by the existence of special-purpose slope detectors in the auditory pathway. However, it is more intuitively acceptable to suppose that inclination has as much effect on surrounding accents as declination, and this is the effect predicted by the model expressed in F18, which has been developed purely on the basis of the scaling of prominence of individual accent configurations.

In Figures 5.19 and 5.20 are F0 contours in which the F0 value on the second peak has been computed according to the model in F18, using pitch prominence factor variation models 3 and 4 respectively. It can be seen that if Model 3 is used, the F0 value of the second peak is increased relative to its value when there is 'baseline' declination between the two accents (as in Figures 5.15 and 5.16); that is, the prominence-F0 ratio is decreased in accordance with the revised version of the LDH. If Model 4 is used, on the other hand, the F0 value of the second peak is reduced relative to its value in the presence of 'baseline' declination, which is contrary to the prediction of the revised LDH. It also happens to be contrary to the prediction of the GDH, which would have the same F0 value on the second peak as in the case of baseline declination.

Again, the same comments could be made about the existence of unnecessary salient points in the contour as were suggested could be made in the case of the falling head (this time in respect of point $C1_2$). In Figures 5.21 and 5.22 are F0 contours in which the prepeak contour element of the second accent acts as a rising head, points $C1_2$ and $C3_1$ being superimposed on points $C0_2$ and $C2_1$. In the case of Model 3, the prediction of the revised LDH is maintained (though perhaps to too great an effect). In the case of Model 4, it is again contradicted, but to an extent which is a little more in keeping with the GDH than in Figure 5.20.

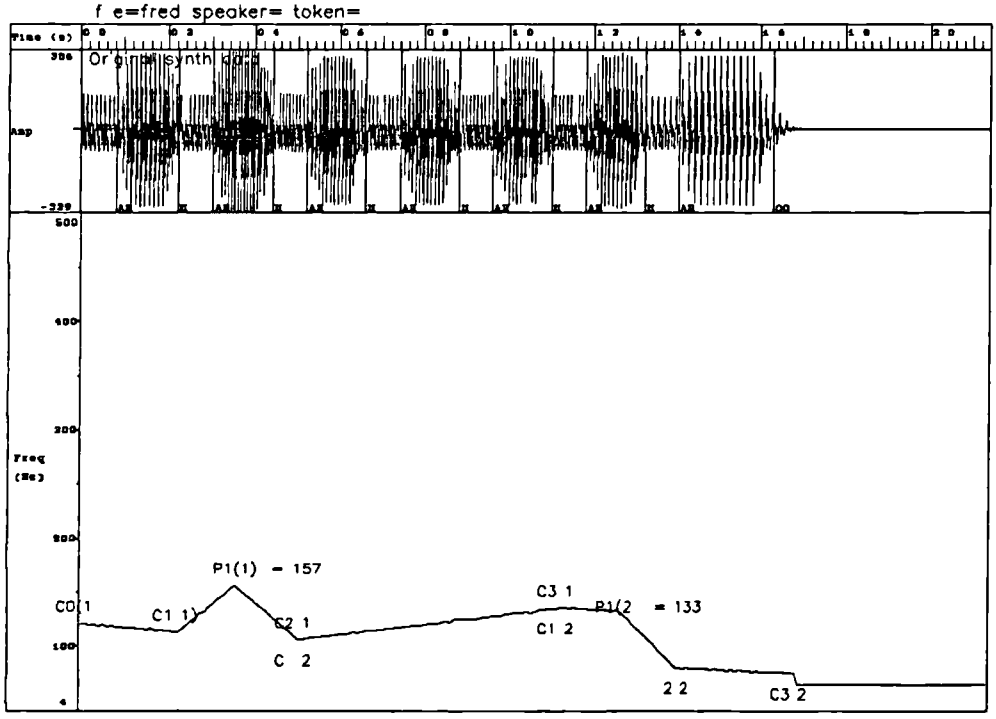


Figure 5.19 F0 contour with rising head, second peak computed using Model 3 in F18.

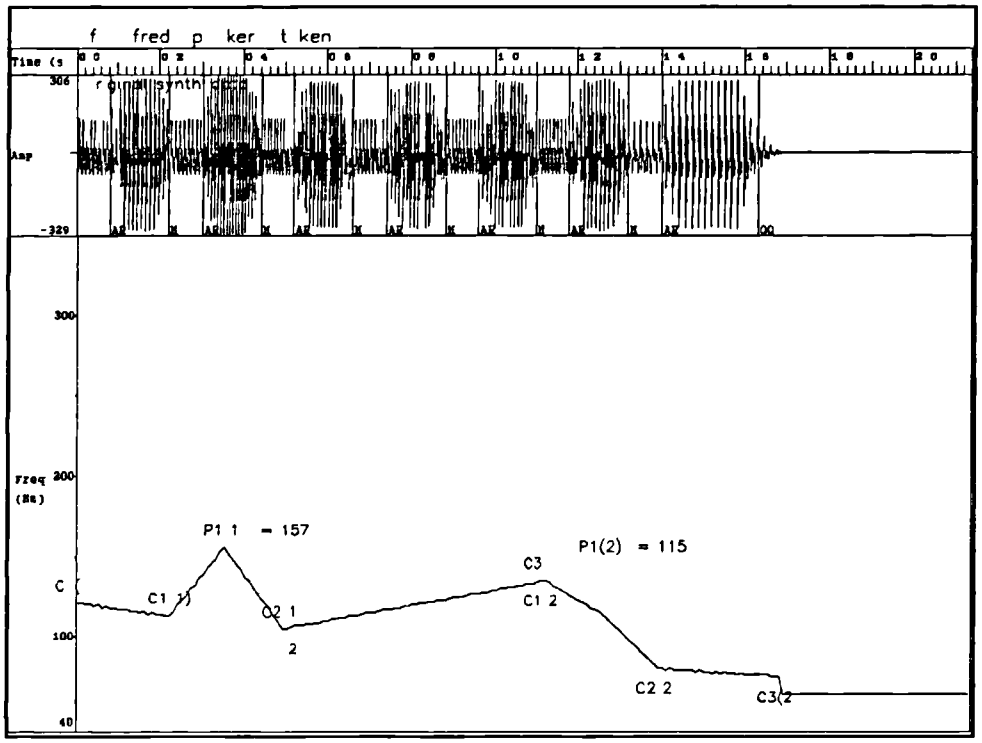


Figure 5.20 F0 contour with rising head. Second peak computed using Model 4 in F18.

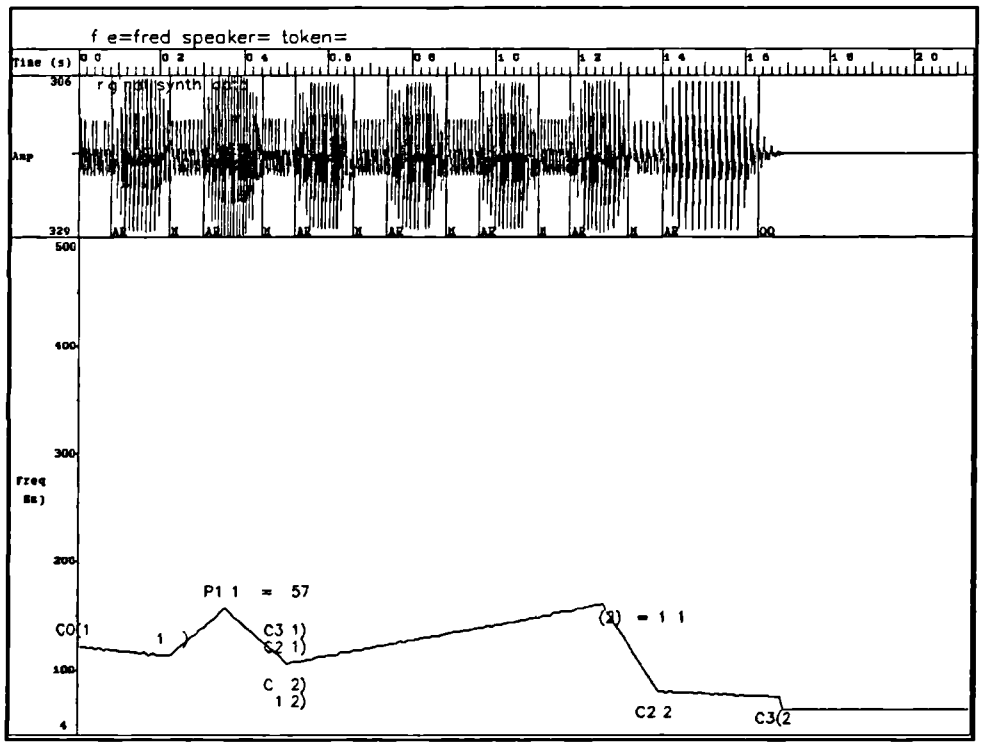


Figure 5.21 F0 contour with prepeak contour element of second accent effectively acting as rising head. Second peak computed using Model 3 in F18.

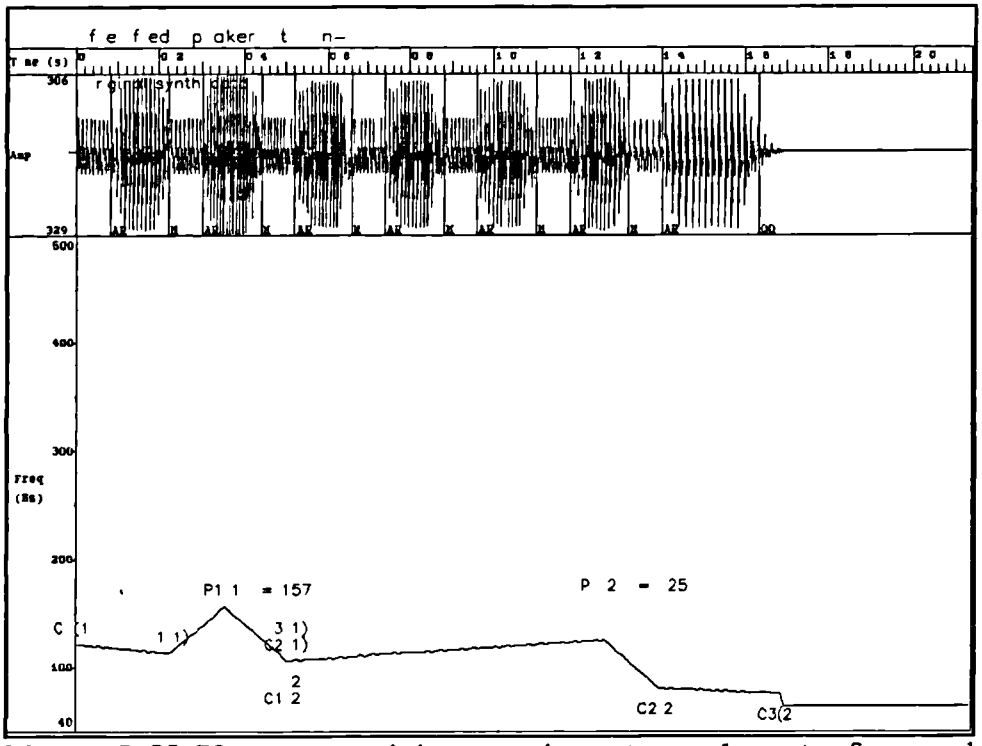


Figure 5.22 F0 contour with prepeak contour element of second accent effectively acting as second peak. Second peak computed using Model 4 in F18.

5.3.6.4 Implications for the Local Declination Hypothesis

The significance for the Local Declination Hypothesis of the maintenance of particular behaviour, even when the theoretical interaccentual contextual component of model in F18 has been eliminated, is that the hypothesis is still valid, regardless of whether a particular contour element has been identified as being the 'vector' for the declination. The model in F18 has different terms for accent and context, and so allows for the possibility of there being some theoretical difference in the behaviour of these terms, but even when the interaccentual context has not been identified as such, but there is still a gradually sloping F0 contour before the accent, a result qualitatively in keeping with the Local Declination Hypothesis is elicited, at least for Model 3. This is a result of the prediction that the pitch prominence function for contour elements should be the same, apart from a peak alignment factor, for accentual contour elements as for contextual contour elements.

At the same time, if the underlying pitch prominence factor variation function does not differentiate between the prominence of contour elements on the basis of slope, as Model 4 effectively does not in respect of elements above

0.167 seconds in duration, then variation of slope in the contour between accents has no effect on the relative prominence of the accent peaks, which is consonant with the Global Declination Hypothesis (if that slope doesn't follow the course of a global declining baseline or topline). The large effect of variation in slope on the pitch prominence of a contour element predicted by Models 1 to 3 results in there being a contingent delineation of contour elements within the model expressed in F18. Those that are steep will tend to be interpreted as accentual by a listener, and used as accentual by a speaker; those that are gradual will tend to be interpreted or used as contextual. The exception to this state of affairs is the class of step accents; only the prepeak contour element of a step accent will tend to be steep, and that is likely not to occur in a perceptually salient position. It is likely that the peak alignment factor, along with a separate amplitude prominence function, is used by speakers and hearers to differentiate between level contour elements in a step accent and level contour elements in the context.

This state of affairs has one particular implication which should be mentioned in passing: that any such contour element could be subdivided into a set of subelements, and prominence functions determined separately for the individual subelements and added together to give a pitch prominence value for the original contour element. How this would precisely be done is not investigated here, but it allows for the possibility of subsidiary accents being defined within the pre- or post-peak context of a main accent.

5.4 PREDICTING PEAK HEIGHT IN DOWNSTEP SEQUENCES

One of the characteristics of the model is that it doesn't make excessive use of unmotivated constants; this is a natural requirement of an approach to modelling the processes of production and perception. The only necessary[1] constants in the model in F18 are the canonical duration of 0.167 seconds and the enhancement constant E. The former appears throughout the model; this is considered justified by the likely involvement of a gating clock with such a period (or similar, or range of period values) in the processes of both the

[1] The use of the base of 1.5 in the sigmoid power function in the second piece of Models 2 and 4 depends on actual use of those models. It was chosen to give an adequately gradual transition from the upper asymptote to the lower. The constant 20.0 identifies the point in time at which that transition is halfway through its course. It could be considered an empirical constant, but really it is only used in this chapter for illustrative purposes.

production and perception of intonation. The enhancement constant is really only present because of a structural requirement of the formula for the pitch prominence of an accent. (Given the considerations here, it has two meaningful values, 1 and 0.667; those values could be made to depend on the existence or otherwise of a partnering accentual contour element to the one to which the constant applies).

Amongst other things, this means that there is no downstep constant in the model. In fact, one of the requirements of this approach is that such an element is unnecessary in a performance model of intonation. This is because of the way the F0 value on the peak of an arbitrary accent is said to be determined both by the peak F0 value of its predecessor and by the shape of it and its predecessor's surrounding context, and notably by the lookahead component that forms that accent's postpeak context.

In the following discussion, it will be assumed that the pitch prominence of a downstepped accent is intended to be the same as that of its predecessor. This is a natural supposition, because downstep is typically involved in episodes of speech in which no demarcative accentuation is performed, such as reading from texts or producing lists, these sometimes bordering on the ritualistic (see Johnson and Grice, 1990). This assumption avoids the problem of the conflation of the effects of prominence and downstep referred to in respect of Pierrehumbert's model in Chapter 3. In fact, prominence and downstep can be seen to be different sorts of categories in a complete model of intonation. Prominence, as a contribution to the F0 values on accent peaks, is more naturally a part of a performance model of intonation, as demonstrated here, and downstep as a process, more naturally a part of a competence model of intonation. This does not prevent downstep as a phenomenon being a natural result of a performance model organised in a particular way, just as it doesn't prevent prominence, as an unanalysed primitive, being a necessary input to a competence model.

5.4.1 The nature of downstep within the developed model

How does a putative process of downstep fit into the scenario presented in Figure 5.12 ? If the F0 value of the first accent in that figure is chosen to be a particular value for the purposes of achieving a certain amount of prominence, on the basis, inter alia, of what contextual contour element

follows the accent, then the peak F0 value of the second accent eliciting that same prominence value, on the basis of, perhaps, the prominence value of the same contextual contour element in the role of prepeak contextual contour element, and the prominence value of the 'lookahead' postpeak contextual contour element, may turn out to be one which is reduced relative to the first by a particular factor. If the same relative configuration occurs in the case of the next accent in the sequence, then to achieve equal prominence, its peak F0 value (or rather, the difference between the peak F0 value and a putative asymptote) may also be reduced relative to its predecessor by the same factor. Thus, a downstepped sequence of peaked accents is initiated.

Yet there is no reason to suppose that there is anything special about such a sequence, such that it warrants the use of a particular mechanism to generate it in a performance model of the kind presented here. It is true that accent sequences tend to preserve the form of the initial accent in the sequence, so that a sequence of peaked accents or a sequence of step accents is more common than an alternating sequence of peaked and step accents (cf. Crystal 1969), and that if this form is copied fairly precisely from accent to accent, and the configuration of the accent is such as to require reduction of the peak F0 value of the order observed in an asymptotically declining sequence to maintain equal prominence, then a sequence may result which could be modelled by the use of a reference asymptote and application of a downstep constant, as in Liberman and Pierrehumbert 1984 (see Chapter 2 for details). On the other hand, if the contexts surrounding the accents are not such as to require effective downstep of peak F0 values for the maintenance of equal prominence, then such sequences won't result. For instance, if there is no identifiable interaccentual context, then the current model predicts that the F0 value of an equal prominence second peak P_{i+1} in a sequence will only be reduced by lower values of $C2_{i+1}$ and $C3_{i+1}$, ceteris paribus. The effect then is not one of downstep, but of postpeak lowering, which in tone-unit final position becomes final lowering. Similarly, if there isn't replication of the accent configurations in a sequence, then the chances are that the sequence of peak F0 values maintaining equal prominence over the accent set won't be asymptotically declining.In short, downstep is not a productive process in speech as far as peaked accents are concerned.

When it comes to step accents, the model in F18 doesn't naturally predict downstep sequences when the main lookahead component used in the model is that of the postpeak contextual contour element. This can be seen from the process illustrated in Figure 5.23.

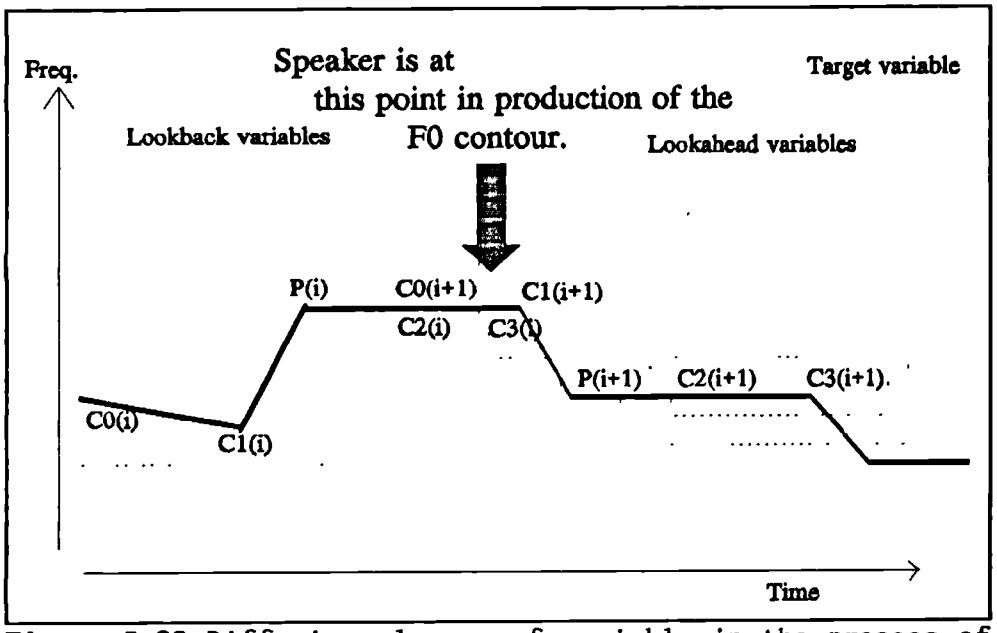


Figure 5.23 Differing classes of variable in the process of producing a downstepping F0 contour with perceptual constraints.

For the correct downstep sequence to be computed in this case, two amendments have to be made to the model. Firstly, the requirement that, for any k, $P_k = \max(P_k, C2_k)$ and $P_k = \max(P_k, C1_k)$, expressed just before F17, has to be lifted. If it is not, then given the values of the lookahead variables $C2_{i+1}$ and $C3_{i+1}$ of the sort demonstrated in the figure, the F0 value on P_{i+1} imparting equal prominence would be higher than that of P_i. This is because of the preaccentual point $C1_i$ being significantly lower than P_i, meaning that the prominence of the initial peak is relatively high, despite the level second accentual contour element. The problem is partly one of the form of the downstep sequence depending on the context at the initiation of that sequence.

Secondly, if a particular F0 value for $C2_{i+1}$ and $C3_{i+1}$ is chosen as the target level of the downstepped step accent, then the model may not compute the correct value for P_{i+1} to match those points, given the configuration of the first accent. Thus, iterative numerical analytic methods would be required to compute a convergent solution to a set of equal values for P_{i+1}, $C2_{i+1}$ and

$C3_{i+1}$, for which the prominence of the i+1th accent configuration matched that of the ith. But the natural interpretation of such a requirement is that the variables $C2_{i+1}$ and $C3_{i+1}$ have ceased to be lookahead variables, but have become joint target variables with P_{i+1}, just within a single model F18[2] (or an amended version of it), whose computation, given that they are all expected to have about the same F0 value, is determined by the requirement to maintain a particular level accent configuration.

Now although it appears appropriate that the amount of lookahead be reduced in downstep sequences, given the contexts in which they are used (reading lists, etc.), it is likely that the revised model wouldn't accord with the correct asymptotically declining sequence[3]. This is because in a step accent sequence, the effect of the ratio max(P,C1)/min(P,C1) in the model of accentual pitch prominence predominates (if a basic pitch prominence factor variation model - 2 or 4 of Models 1-4 discussed in 5.2.6 - is chosen such that variation in the duration of a level contour doesn't result in much variation in pitch prominence), since the effects of the pre- and post-peak contextual contour elements cancel each other out, and the other accentual ratio P/C2 is always 1 in the step configuration. This means that the major determinant of the F0 value at which the equal prominence solution for the peak and postpeak contour converges would be precisely that former ratio (which, in the step accent, becomes C1/P). In an initial configuration of the sort depicted in Figure 5.23, the solution will have the value of P_{i+1} (and the following variables $C2_{i+1}$ and $C3_{i+1}$) a sizeable distance below the value of P_i, because of the size of the ratio $P_i/C1_i$. Subsequent peaks are bound to show an asymptotic decline, because, for any k, the value of $C1_k$-P_k required to maintain the ratio $C1_k/P_k$ is bound to decrease as P_k decreases.

Because of the predominance of the factor C1/P in determing the prominence of the step accents in the sequence, the value of P_{k+1} can be roughly approximated using the series

$$P_{k+1} = P_k/(C1_k/P_k) = P_k^2/C1_k$$

(where k>1, since P_2 has to be computed by numerical analysis).

---

[2] That is, not joint target variables in the network of F18-type models suggested to be operative in the production of intonation contours.

[3] It should be said that the full revised model suggested was not implemented for this thesis.

Since on each step, the value of $C1_{k+1} = P_k$, the value of the k+nth term of that series is given by

$$P_{k+n} = P_{k+1}^{k+n}/P_k^{k+n-1}$$

It can be seen that this means that for typical initial step values (between P_1 and P_2), the subsequent downstep sequence would go too low, fairly rapidly. For example, if the F0 value of P_1 were 157Hz and the F0 value of P_{1+1} 134Hz, P_{1+4} = 83Hz, which is clearly too low. Therefore, unless other factors are made to compensate, the model incorporating lookback and lookahead variables from adjacent accents seems inappropriate for an account of the production of downstep sequences.

5.4.2 The nature of downstep in a model using longer lookahead

The amendments to the model which would at minimum be required to make downstep sequences emerge naturally from the model in F18 need to be made in any case. That is, for two step accents i and i+1, where accent i is the first in a sequence of such accents, the predicted F0 values on the points P_{i+1}, $C2_{i+1}$ and $C3_{i+1}$ should be computed by numerical analysis[4]; and some mechanism should be introduced such that the value returned by the function can be determined either to be a peak or trough point, so that an appropriate denotation can be given to it, given a target configuration.

If it is assumed that such amendments have been made, an alternative interpretation of how downstep sequences could result can be presented. This is schematised in Figure 5.24 .

[4] This could be done even if the F0 values on those points was not intended to be equal; for instance, they could form a rising configuration, which would then be repeated on subsequent accents in the following sequence.

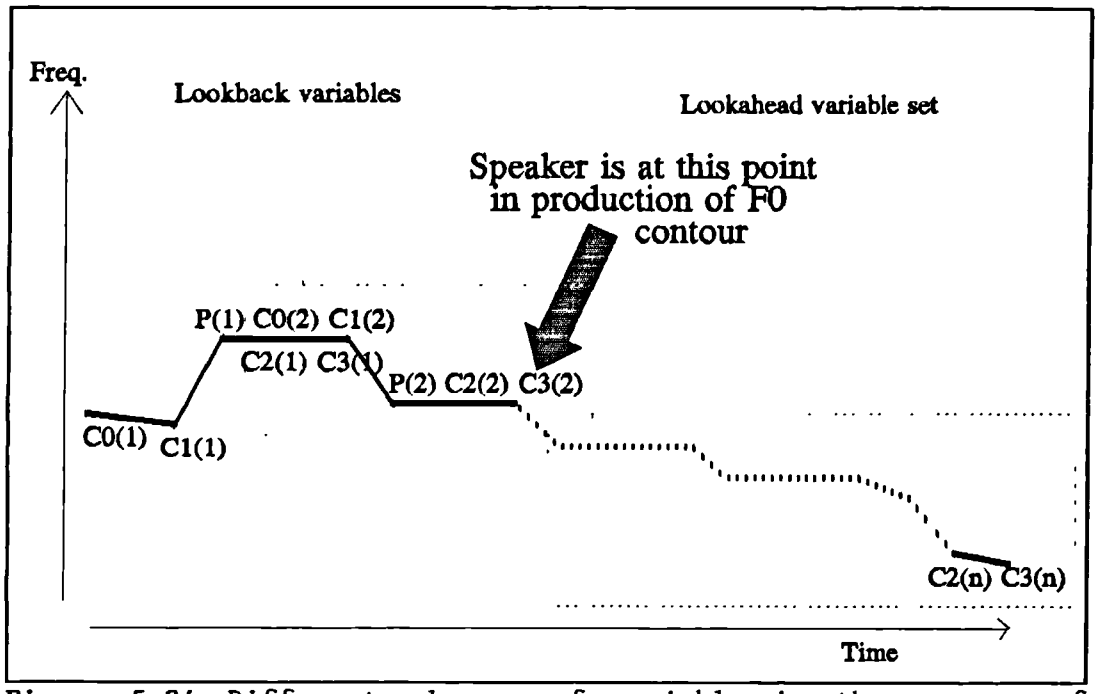


Figure 5.24 Different classes of variable in the process of producing a downstepping sequence, in a model with long lookahead.

In this figure, it is proposed that instead of the pre and postpeak context of an imminent accent forming the lookahead variables in the production of a downstep sequence, instead it is the prepeak context of the imminent accent and the postpeak context of the final accent which have that function. This arrangement takes advantage of the fact that the final fall in a declarative utterance tends to reach a pretty constant value for any one speaker under particular conditions; the speaker will be assumed to have accessible as lookahead variables rough estimates of the F0 values of the salient points in the postpeak context of the final accent, because that information does not vary much.

Formally, this change in the postpeak lookahead variables of the imminent accent results in a revised version of the model in F18, here shown as F19:

$$F19:\ P_{i+1} = C1_{i+1} * C2_n \,/\, (E * (C2_n * (F(C1,P) * T(C1_{i+1}, P_{i+1}) + A_{c1})) + C1_{i+1} * (F(C2,P) * T(C2_n, P_{i+1}) + A_{c2}))) * (J * (\max(C1_i, P_i)/\min(C1_i, P_i) * E * (F(C1_i, P_i) * T(C1_i, P_i) + A_{c1}) + \max(P_i, C2_i)/\min(P_i, C2_i) * E * (F(P_i, C2_i) * T(P_i, C2_i) + A_{c2}) + C1_i/C0_i * F(C1_i, C0_i) * T(C1_i, C0_i) * R_{c1,c0} + C2_i/C3_i) * F(C2_i, C3_i) * T(C2_i, C3_i) * R_{c2,c3}) - C1_{i+1}/C0_{i+1} * F(C1_{i+1}, C0_{i+1}) * T(C1_{i+1}, C0_{i+1}) * R_{c1,c0} - C2_n/C3_n * F(C2_n, C3_n) * T(C2_n, C3_n) * R_{c2,c3})$$

(for i=1 to n-1, where n=number of accents in the tone unit).

If this scenario is used as the basis for the production of downstep sequences, the sort of contour in Figure 5.25 (using Model 3 in F19) is generated on a stretch of reiterant speech. In the model, unadjusted in respect of the requirement to use numerical analysis for the computing the height of P_2 and its following context, an appropriate such value has been computed anyway. It should be remembered, though, that the computation of this value is sensitive to the values of $C0_1$ and $C1_1$. A similar sequence is computed using Model 4 (the final peak is at 110Hz in this case).

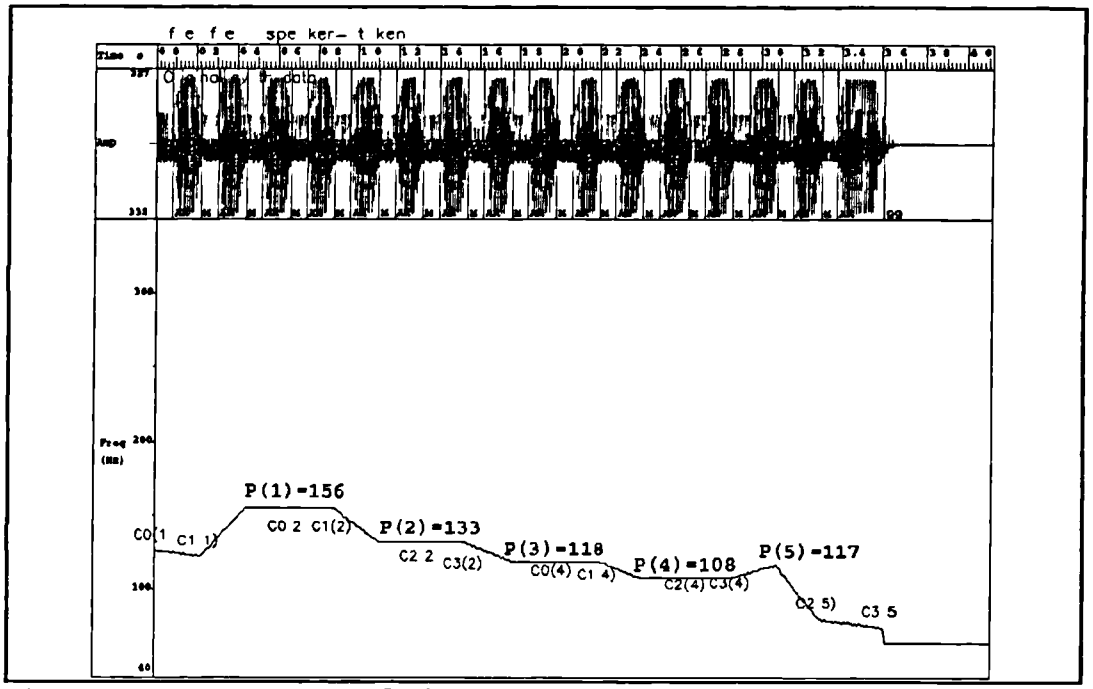


Figure 5.25 Sequence of downstepped step accents computed using long-lookahead model in F19 (Model 3). Identifiers of some points have been removed to improve legibility.

The use of a default sort of lookahead is appropriate in a mechanism for producing downstep sequences in intonation, because of the contexts in which downstepping sequences of the sort exemplified tend to be produced (reading from texts, producing lists, etc). The strategy of the speaker is interpreted as being to use what little information is available in an available postpeak lookahead pool to produce an acceptable contour, and that information is the 'tail' they would expect to produce after the final accent in a declarative utterance. Downstep is thus viewed as involving the invocation of a global perceptually-constrained mechanism in the production of intonation in order not to assign differential prominence to peaks in a sequence.

On the other hand, when an individual accent, be it stepped or peaked, occurs lower in pitch than its predecessor, the strategy involved is interpreted as being one either of adjustment of the peak F0 to the known preaccentual and predicted postaccentual context, or adjustment of the said contexts to a targetted peak F0 value, in order to produce the desired effect of prominence relative to the predecessor (be it equal or different). The

latter local perceptually-constrained mechanism would be expected to be more often used in spontaneous speech.

5.5 THE FORM OF F0 DECLINATION

Within the accent-by-accent strategy, proposed at the end of the last section to be involved in spontaneous speech, are involved two types of more general mechanism, already mentioned in section 5.3.5 above. In one, the production of intonation patterns is more perceptually constrained than in the other:

Mechanism Type 1 (auditory control of context)

In this mechanism, an example of which is depicted in Figure 5.12, the lookback variables of the immediately preceding accent are used, along with the lookahead variables of the upcoming contextual contour elements, in a model which is used to generate a reference slope for an interaccentual stretch, given a target pitch prominence value for the upcoming accent configuration. This is used in a feedback mechanism in which that part of the auditory input corresponding to the F0 signal is tested for its slope, and appropriate adjustments made to the motor control mechanisms of the expiratory and inspiratory respiratory muscles, and the intrinsic and extrinsic laryngeal musculature, to compensate for any deviations from the reference slope.

In the other, there is more relinquishment to passive mechanisms involved in the production of intonation contours:

Mechanism Type 2 (passive determination of context)

In this mechanism, the F0 signal is not tested for slope in the context, but the start and end points of contextual contour elements are accessed, so that the accentuation mechanism can use the context against which the height of the upcoming accent peak must be scaled to convey a certain value of pitch prominence. The role of auditory feedback in controlling the slope is reduced, and possibly eliminated, so that kinaesthetic feedback mechanisms predominate in maintaining the character and continuity of phonation, and passive mechanical aspects of the physiological machinery involved in producing intonation contours predominate in determining the form of contextual contour elements.

It is not suggested that these two types of mechanism are mutually exclusive, either in theory or practice. That is, both mechanisms are suggested to occur to greater or lesser degrees in the production of intonation, and could possibly both occur in the production of any particular intonation contour, especially in spontaneous speech. However, if they are considered separately and exclusively, then some conclusions can be drawn about the form of declination after the involvement of subglottal pressure variation is taken into consideration.

The task of the mutual adjustment of the respiratory and laryngeal musculature to maintain a targetted interaccentual slope is simplified if the adjustments that are made are restricted to one or the other of the sets of muscles. Thus it is suggested here that the overall strategy in maintaining that slope would be to maintain a constant short-time averaged transglottal pressure level[5], such that increase or decrease of vocal-fold tension effected by adjustments of the laryngeal muscles, or planned relaxation of vocal-fold tensors (cf. Titze and Durham 1987), becomes the sole determiner of the course of the frequency of vocal fold vibration (apart from supralaryngeal articulation). Since transgottal pressure varies as vocal-fold tension varies, ceteris paribus, this would also require the use of tactile and proprioceptive feedback loops (as is suggested exist, using mechanoreceptors in the subglottic mucosa and proprioceptors in the laryngeal muscles and articulations of the laryngeal cartileges [Wyke 1974, Adzaku and Wyke 1979, Adzaku and Wyke 1982]) to independently maintain a constant subglottal pressure head. Thus, a mechanism of type 1 would have auditory feedback control of the slope of declination (or inclination) in unaccented stretches of speech and laryngeal (and probably respiratory) feedback control of the maintenance of adequately non-varying transglottal pressure.

If there is no targetted interaccentual F0 slope, but the speaker just responds, in the form and degree of accentuation, to the slope that happens to be produced by allowing passive mechanical forces to predominate, then there is no requirement to maintain a constant transglottal pressure. Indeed,

---

[5] This is suggested as a general strategy in speech by Ladefoged 1963, compromised only by the interference of supraglottal articulations, which thereby also compromise targetted intonation contours of the form exemplified in this chapter.

the transglottal pressure will then tend to decline during the interaccentual slope (if constant airflow out of the lungs is assumed) because of the continual reduction in lung volume, the effect of which has to be counteracted to maintain constant transglottal pressure. Thus, a mechanism of type 2 would not necessarily have any specialised feedback involved in the maintenance either of a particular slope of declination or of a particular constant value of transglottal pressure. However, there could be involved more general mechanisms of auditory and laryngeal feedback to ensure that neither the frequency of vocal-fold vibration nor lung volume decreased too rapidly for the purposes of an intended utterance.

The expected course of transglottal pressure and corresponding course of F0 for the two mechanisms is thus as in Figure 5.26. The discussion so far has been in terms of the course of F0 during contextual contour elements, and the local F0 configurations displayed in that figure are intended as representations of the kinds of contextual contour elements that occur during the operation of the two respective mechanisms[6]. But what can be said of the behaviour and control of accentual contour elements under the same mechanisms?

Well, the discussion in subsection 5.3.6.4 points out that there is a contingent delineation between contour elements interprable as accents and those interprable as context by certain prominence factor variation models (notably Model 3), on the basis of the slope of the contour element. It is likely this contingent delineation is reflected by differing susceptibility to the form of control suggested in Mechanism 1. This is because for steep F0 movements, the role of auditory feedback is likely to be greatly diminished, because the durations of the accentual movements of, say an octave or less, are too rapid for feedback to be effective.

---

[6] It has to be remembered firstly, that the types of mechanism could occur at different times within the production of a single intonation contour, and secondly, that the effects of supraglottal articulations on the course of F0 are considered to have been factored out of the contours.

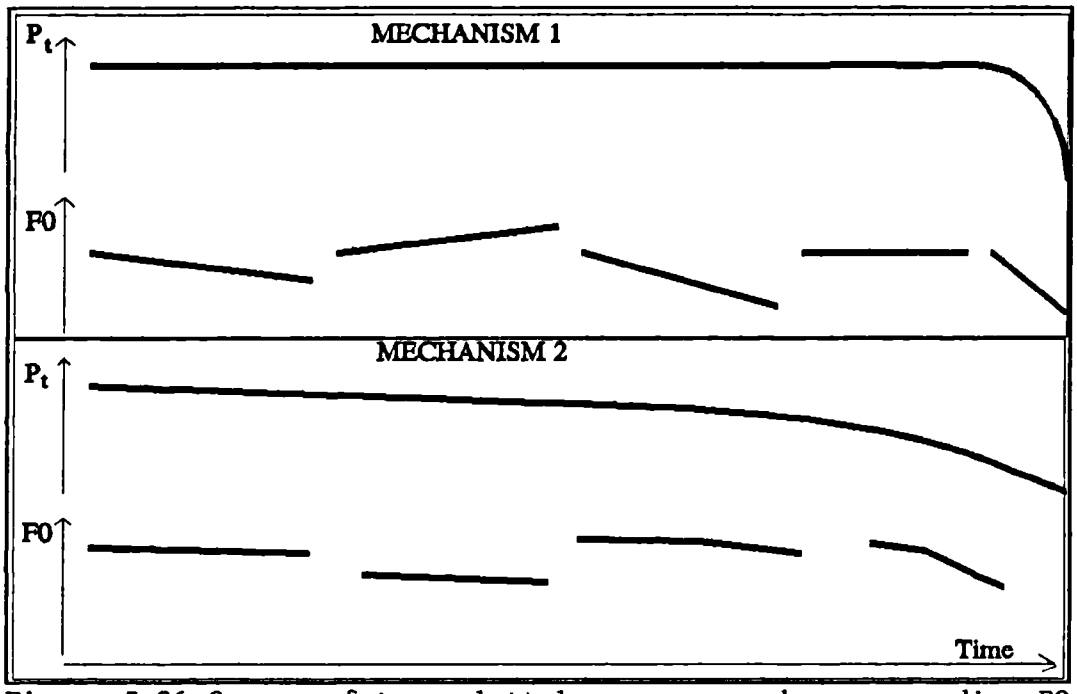


Figure 5.26 Course of transglottal pressure and corresponding F0 expected under the two mechanisms of the production of intonation during contextual contour elements, discussed in the text.

This does not prevent both accentual and contextual contour elements being responsible for the declination effect, again as discussed in subsection 5.3.6.4. If accentual movements are controlled by a mechanism similar to Type 2 insofar as the F0 value is only auditorily monitored, rather than auditorily controlled, then it would follow that the mediation of the declination effect is independent of the use of a feedback mechanism for controlling the slope of a local F0 contour prior to the accent on which the declination effect takes place. Both the Local and the Global Declination hypotheses are thus consistent with either of the types of mechanism suggested to be involved in the production of intonation.

The Local Declination Hypothesis predicts that the declination effect in production is mediated only by the tendency for declining contextual contour elements to occur and be adjusted for or compensated for in the production of a following accent peak for a desired degree of prominence. It predicts that the declination effect in perception is mediated by the existence of contextual contour elements that tend to be declining when surrounding accentual contour elements in a total accent configuration. But the

declination effect in general is not a necessary accompaniment to the production or perception of an arbitrary intonation contour.

The Global Declination Hypothesis predicts that the declination effect in production is mediated either by the existence of declining transglottal pressure, or perhaps by a perceptual constraint that a cued (or uncued) baseline and/or topline be followed during generation of an intonation pattern. It predicts that the declination effect in perception is mediated by the scaling of accent peak within a declining frame of reference.

From this discussion, it can be seen that the form of declination is an object of study that properly comes within a competence model of intonation. This is partly due to the reification of the concept since its introduction by the Eindhoven School in the 1960's, such that it has become almost an object to be systematically manipulated, and its characteristics to be investigated, without regard to questions of whether it is an appropriate concept. More meaningful observations about its form can be stated under the Global Declination Hypothesis.

Under that hypothesis, a competence model of intonation has a declining frame of reference for the scaling of intonation contours (always, in the case of the Eindhoven School, and sometimes in the case of an analysis such as Thorsen's[7]) and in a performance model uses the same frame of reference as control in production and for decoding in perception.

The Local Declination Hypothesis is apparently tied to a performance model of intonation (though it is equivocal between which of Mechanism Types 1 and 2 is used in the production of F0 contours). Under it, the form of declination in general can be seen to be illustrated only by the declining types of contextual contour element appearing in Figure 5.26. For incorporation within a standard competence model, a translation would be required of the model presented in this chapter, which is based on pitch movements, into pitch level primitives, which are more abstract entities. In this case, 'proper' statements about the form of declination under the Local Declination Hypothesis could be made. But in that case, the contour elements most likely

---

[7] For both, see Chapter 2.

to be the exponents of the Local Declination Hypothesis would be relegated to the status of interpolations between F0 targets, and the form of declination would again have to be described in terms of a global frame of reference.

However, a synthesis of the positions expressed in the Local and Global Declination Hypotheses can be made if it is considered (a) that the Local Declination Hypothesis is part of the description of the operation of the intonation production and perception processes at a lower (subcortical) level of the Central Nervous System, whereas the Global Declination Hypothesis is part of those processes at a higher (cortical) level, dealing with more abstract processes, and/or (b) that the Local Declination Hypothesis is part of the description of such processes in the developing child, whereas the Global Declination Hypothesis describes the operation of such processes in a human beyond a certain stage of development. If the latter suggestion were true, and it seems plausible, the only inconsistency in the suggested scenario would be that the Local Declination Hypothesis, part of a peformance model of intonation, would then form a part of a description of the neonate's and infant's developing intonational machinery, which is the archetypical domain of a competence model of intonation.

Perhaps, then, further investigation of these issues would reveal that intonational competence is a facility that ascends from the CNS periphery to the neocortex during development, and that intonational performance consists of higher-level adaptation to the developing peripheral processes which, still in place, have the effect of modulating the higher-level process during adult life.

In the following chapter, the question of one aspect of this latter interaction between lower and higher levels of the Central Nervous System is implicit in an investigation of the existence of auditory feedback in the control of intonation.

CHAPTER 6
THE AUDITORY CONTROL OF DECLINATION

6.1 INTRODUCTION

Towards the end of the previous chapter, two different types of mechanism were suggested to be operative in the production of intonation contours. In one, an auditory feedback loop is used to control the slope of individual contour elements, where a contour element is defined as the stretch of F0 from one salient F0 value to the next. It is theoretically possible that such control be exerted over both accentual and contextual contour elements, though a contingent delineation occurs between those elements on the basis of slope (as well as a modelled delineation on the basis of alignment to syllable structure), such that it is more likely that contour elements of gradual slope, which tend to be contextual contour elements, have their slope controlled in this way than those of steep slope, which tend to be accentual contour elements.

In the other type of mechanism, the slope is not a controlled variable in an auditory feedback loop; instead, the slope that arises as a function of the interaction between passive mechanical forces and muscular control without the fine adjustments required of a feedback mechanism to effect a particular slope in the change in the rate of vocal fold vibration (such as might occur in the passage from one physiologically controlled target to the next) is monitored, and the value that it takes used as input to control mechanisms involved in adjusting the configuration of later parts of the intonation pattern.

The coexistence of these two mechanisms in the production of intonation implies the possibility of a range of intricate auditory feedback mechanisms in its control. To test for them would require a long sequence of individual experiments, investigating different aspects of the mechanisms proposed. In particular, for the the first mechanism proposed, a sophisticated set of tests is required involving the adjustment of a fedback F0 signal, appropriately filtered to account for the differential effects of bone and air conduction in such a way that the type of disruption, to the signal that would normally be detected via these two media, can be predicted and thus controlled.

There are other feedback mechanisms in which the nature of the control would appear to be simpler. Both mechanisms, in fact, allow for the existence of control for the maintenance of the absolute slope of change in the rate of vocal fold vibration above a threshold value. This control would typically be required during local stretches of declining Fx[1], so that vocalisation did not become impossible due to the degree of vocal fold abduction or to the thickness of the vocal folds due to relaxation of the vocal fold tensor muscles, and so that there was not an inappropriate switch into creak[2].

Therefore, the first experiments in the sequence exploring the auditory control of declination should test for the mere existence of auditory feedback during unaccented speech, since that is the prerequisite to the maintenance of the absolute slope above threshold. For stretches of unaccented F0, the expectation would be that if the possibility of auditory feedback is removed, then more instances would occur of the patterns of Fx during those unaccented stretches containing occasional steep descents, thereby transgressing a putative threshold slope. However, it is unclear what such a threshold slope should be in any particular circumstance, and so the very initial experiment should be an exploratory one, to see how the typical slope used on interaccentual stretches for a particular sort of intonation contour changes in the absence of auditory feedback, and to see whether the occasional steep descents occur.

At the same time, there are differential predictions about the existence of auditory feedback in intonation depending on how strong the distinction is made between accent and context. Accents and context are contingently differentiated not just on the basis of slope, as was discussed above, but also on the basis of relative amplitude. Accented syllables have in general significantly higher amplitude than unaccented ones. There is thus a twofold justification, on the basis of observable differences between accentual and contextual contour elements, for envisaging that either auditory feedback is

---

[1] See chapter 1 for an introduction of the quantity Fx.

[2] This control is often relinquished at the end of an utterance, during final lowering, so that creak results (see Figure 2.11 in Chapter 2).

involved in only one of the types of pitch movement[3], or that if it is involved in both, the mechanism is different, possibly involving feedback loops at different levels of the ascending auditory pathway.

For these reasons, an experiment was devised in which the aim was to disrupt auditory feedback continually, and then differentially between accent and context, in the production of an intonation contour containing two equally configured accent peaks, and to observe the respective effect on the peaks and troughs of the intonation contour. The form of disruption chosen was high- amplitude masking noise administered during the course of the utterance, and triggered by features of the intonation contour being produced in real-time.

6.2 NOISE-MASKING AUDITORY FEEDBACK DISRUPTION EXPERIMENT

6.2.1 Method

6.2.1.1 Subjects and experimental set-up

Three male speakers of English, members of the Phonetics Department at UCL, were used as subjects in the experiment. For each subject, the experiment was performed in a sound-treated room. The basic task of the subject was to speak (via a clip-on microphone) under differing conditions of headphone-administered masking noise[4]. The speech that they produced was fed to a computer program via the A-to-D peripherals of a minicomputer, and monitored for amplitude; the Tx[5] they produced was fed to the same program (via a Laryngograph board [Fourcin and Abberton 1971] adjusted

---

[3] The difference in salience due to amplitude would argue for auditory feedback only being effective in the case of accentual contour elements. The difference in operability of feedback due to the time available would argue for auditory feedback only being effective in the case of contextual contour elements.

[4] The signals were administered to the subjects as they spoke via open-cup Sennheiser headphones of 32kOhm impedance. Closed cup headphones are inappropriate for this type of experiment, as they have the effect of amplifying the bone-conducted voice signal of the speaker.

[5] Tx is the reciprocal of Fx, and comprises vocal fold vibration markers determined from the Lx signal.

to generate a pulse train detectable by the minicomputer A-to-D), and monitored for transgression of a threshold level of corresponding Fx, determined for each speaker, and acting as an accentuation threshold. The monitoring program[6] also controlled the administering of three types of signal via the D-to-A peripherals of the minicomputer: silence, a white noise low-pass filtered at 500Hz[7], and the fedback voice of the speaker (see Experimental Protocol). Signals thus administered had line-noise removed from them by a passive attenuator - power amp - attenuator series between the line output and the headphones[8].

The level of experimentally-administered noise was checked to be within acceptable levels (cf. Dixon-Ward 1970) by fitting one headphone cup to a Bruel and Kjaer artificial ear, and noting the calibrated long-time averaged level of a signal which speakers felt, on the basis of a continuously administered such signal, was capable of masking their speech, on an Onisokki Spectrum Analyzer.

6.2.1.2 Recorded data

Three different channels of data needed to be recorded - the speech signal, the Lx signal and, in the case of the administration of a disruptive signal, the signal administered over the headphones. Since a multi-channel recording device was not available, the speech signal was recorded twice, once on audio tape, and once on Betamax video tape. Alongside the speech on audio tape was recorded the Lx signal, and alongside the speech on video tape was recorded the administered noise signal, in those cases when such

---

[6] The program was adapted by the author from an original Tx acquisition program written by Mark Huckvale of the Phonetics Dept. at University College London.

[7] Low-pass filtered noise with that cut-off frequency was considered appropriate for the masking of the voice of a male speaker at high amplitudes. For two of the subjects, this was successful (according to their reports) whereas for the third, high frequency harmonics (resulting most likely from the sharpness of glottal closure in his excitation signal) detectable above the level of noise enabled him to determine the variation in voice pitch, during the running speech stimuli.

[8] I am grateful to Mahen Goonewardene, David Cushing and Richard Baker of the Dept. of Phonetics, UCL, for assistance and advice in setting up the current experiment.

a signal was administered. Alignment of these signals was performed automatically in computer analysis using a cross-correlation routine[9].

6.2.1.3 Experimental protocol

The most basic hypothesis tested in the experiment is that auditory feedback is used in the control of the frequency of vocal fold vibration. This most basic hypothesis can be tested using the production of the most simple token, a sustained open back unrounded vowel ("ah"). In addition, the hypothesis about feedback being used to maintain the slope in the rate of change of vocal fold vibration above a particular threshold level is inevitably tested using such a token, since the threshold slope cannot be anything other than zero.

Confirmation of that hypothesis in what could be seen as a trivial case could not be extended to the case of the production of an intonation contour without testing of that case; but in any case, the intended experiment in the case of the latter is in the first instance an exploratory one on the differential effects on accents and context of masked auditory feedback. This can be performed by having subjects produce a single utterance containing two peaked accents, and varying the conditions of masking noise. Since, for the investigation of unaccented intonation, the variation in slope of the interaccentual F0 contour under the differing conditions is the object of study, an additional parameter can be varied, viz. the interaccentual duration, to see what influence it has on that variation.

There are thus three basic sets of utterance, and four conditions of masking were chosen for the utterances to be spoken in (a cross in the following table marks the testing of the utterance type with the masking condition) :

[9] The success of alignment was checked for each token recorded in this dual fashion. Those for which alignment was unsuccessful were excluded from later analysis.

Utt'nce Type	Masking Condition: No Masking	Continuous Masking	Masking only during unaccented speech.	Masking during unaccented spch; speechfeedback during accented speech.
Sustained vowel	X	X		
Short 2-peaked sentence	X	X	X	X
Long 2-peaked sentence	X	X	X	X

The reason that the fourth condition of masking noise has speech feedback during non-masked (accented) parts of the utterance is that it was feared that one possible vitiation of the experiment would be the tendency for speakers to 'speak up' during episodes of masking noise, in order to provide themselves with auditory feedback. The addition of this sporadic 'sidetone' feedback was intended to compensate for the existence of the masking noise, since "sidetone feedback has a continuously variable effect on voice level, the inverse of the effect of noise" (Lane and Tranel, 1971). As additional counteraction to this vitiating factor, the subjects would be presented with a noise-level meter during the experiment, with an upper level they should try to avoid transgressing.

The two sentences used were

S1: A WILLow may be rarer than a YEW. (short sentence)

and

S2: A MALLow may be rarer even than a WILLow. (long sentence)

with the intended accented syllables marked in capitals. There was no control on the segmental content of the accented syllables, because the subject of investigation was the differential behaviour in intonation during the different conditions of masking feedback.

The intended intonation contour was a double peaked-hat intonation contour. To facilitate the production of this contour, a context was provided for each of the sentences which it was considered would encourage its use (see the Appendix to this chapter for the text of this context). In addition, several renditions of the targetted intonation contour type were made by the author before the actual performance of the relevant part of the experiment, so that the subject had a direct auditory impression of the contour intended.

It was intended that the duration of administration of the high-amplitude masking noise be reduced to a minimum. At the same time, the requirement for real-time triggering of the administration of the noise by events in the intonation of the speaker meant that the noise could be turned on and off by the existence or otherwise of phonation. Thus, noise was only administered when the speaker was speaking. This had the added advantage of avoiding the possibility of habituation to the level of noise.

6.2.1.4 Experimental Procedure

Subjects were asked to read the instructions in the appendix to this chapter. They were then fitted with the laryngograph electrodes (whose satisfactory placement was checked by the display of Lx on an oscilloscope), the clip-on microphone (whose satisfactory placement was checked by the recording of data onto the computer and display of the amplitude level) and the headphones. They were seated in front of a computer terminal connected to the minicomputer, and presented with a menu system which they used to guide them through the experiment. For each utterance type and operative masking condition, they were asked to produce ten repetitions.

There were two stages in the experiment; an initial aquisition stage, in which data in the 'no masking' condition was acquired, and parameters from the acquired speech data to be used in the disruptive masking were determined, and a masking test phase, in which data in the other three conditions

involving noise masking were acquired (onto analog media) and concurrently used to trigger the administration of the masking noise, under the constraints determined by the parameters established in the first stage.

6.2.1.4.1 First Stage (no noise masking)

In the first test, they were asked to produce a sustained "ah" vowel for ten seconds, or for the maximum duration they could sustain it, whichever was the smaller. An online display counted down ten seconds for them once they had started the utterance. While they were speaking, the controlling computer program[10] stored and averaged the short-time rms amplitude of their speech over 40ms windows. The final averaged value for that token was stored alongside the speech, Tx and Fx data. This procedure for storing mean rms amplitude data was repeated for all the tests in the first stage.

In the second test, they were asked to produce the first (short) sentence with the requisite intonation. In addition to the storage of amplitude parameters while this procedure was continuing, an F0 value was determined from the F0 contour for each token, once they had all been acquired, using the following procedure:

The F0 target identification procedure of Hirst and Espesser (Hirst and Espesser 1991) was used to determine the trough points immediately either side of each accent[11]. The average of these four points was stored against each token. For each sentence, the mean of the ten average values stored, plus an offset of 15Hz was determined as the threshold value to be used in triggering the offset of noise masking for this sentence, for those conditions requiring it, during the second stage of the experiment.

---

[10] This was a separate program written by the author which did not provide the facility for feedback, but allowed for the concurrent acquisition of speech and Tx data.

[11] In an informal pilot test prior to running the experiment, the author had confirmed that this method was robust in determining those trough points for the type of intonation contour being elicited. The only problems arose in one or two tokens in which the program containing the algorithm crashed (for unidentified reasons, probably due to implementational problems). In those cases, the token was simply ignored in computing the mean F0 threshold value for that condition.

In the third test, they were asked to produce the second (long) sentence with the requisite intonation. The amplitude and F0 threshold parameters were stored in the same way as for the short sentences.

Before the second stage was performed, the absolute level of administered noise was checked to be within acceptable levels, as mentioned above. In all cases, the long time averaged level of administered noise did not exceed 108dB SPL.

6.2.1.4.2 Second stage (noise masking)

In this stage, noise masking of different forms was administered in the production of the utterances. For the sustained vowel, only one condition of noise-masking was meaningful, viz, continuous noise-masking. At the same time as the subject again uttered the sustained vowel of ten seconds duration, the low-pass filtered noise was administered over the headphones. It was triggered by the onset of phonation, detected by the laryngograph electrodes connected to the pulse train generation card connected to the A-to-D input of the minicomputer. A concurrent sound level meter, with an upper threshold determined from the maximum of the mean rms amplitude values stored in stage 1, and displaying the current level of the voice in real-time, was presented to the subjects on the computer screen. For two of the subjects, if they exceeded the upper threshold level, the computer produced a bleep.

For the sentences, subjects had the three different sorts of masking listed above administered to them in three respective sets of ten utterances. The continuous noise was again triggered by the onset of phonation. For both conditions of the noise which cut out during the accented syllables, the cut-out was triggered on and off by transgression of the F0 threshold determined in Stage 1. In the second such condition, the speaker's own voice was fed back during the cut-out period.

The whole experiment took about an hour for each subject.

Following the experiment, those data which were not already stored on computer disk were acquired onto the computer from the analog media, and analysis performed on them.

6.2.2 Analysis

6.2.2.1 Sustained vowels

Examples of some of the data recorded during the two different conditions for the sustained vowels for two of the three speakers appear in Figures 6.1-6.4.

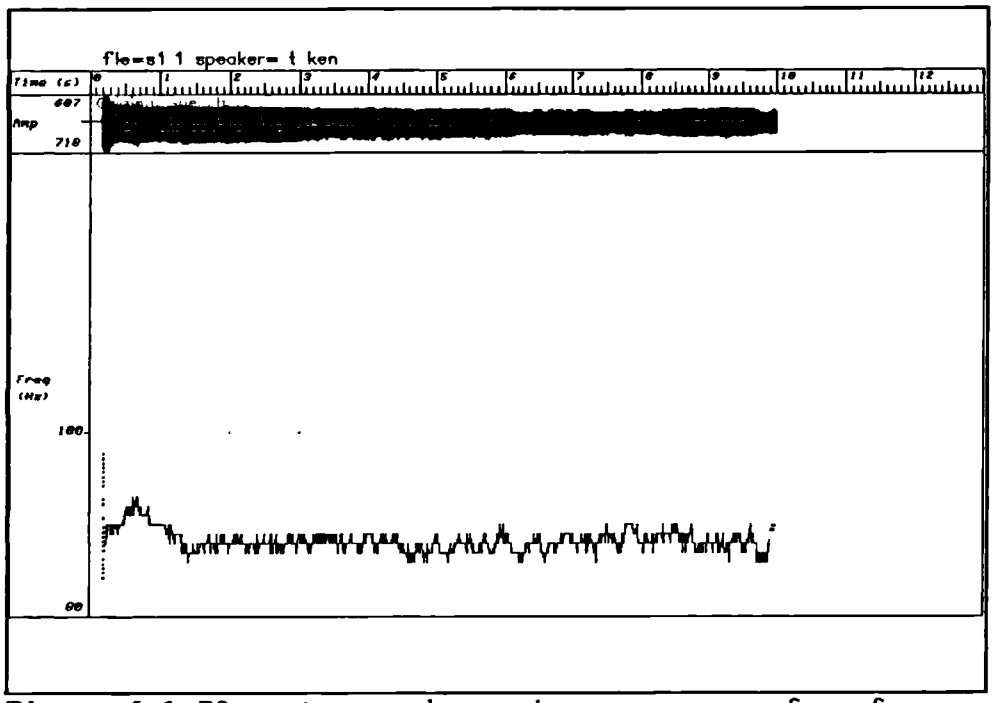


Figure 6.1 F0 contour and speech pressure waveform for one instance of the sustained vowel spoken by speaker A, without masking noise. F0 axis: 80-130Hz

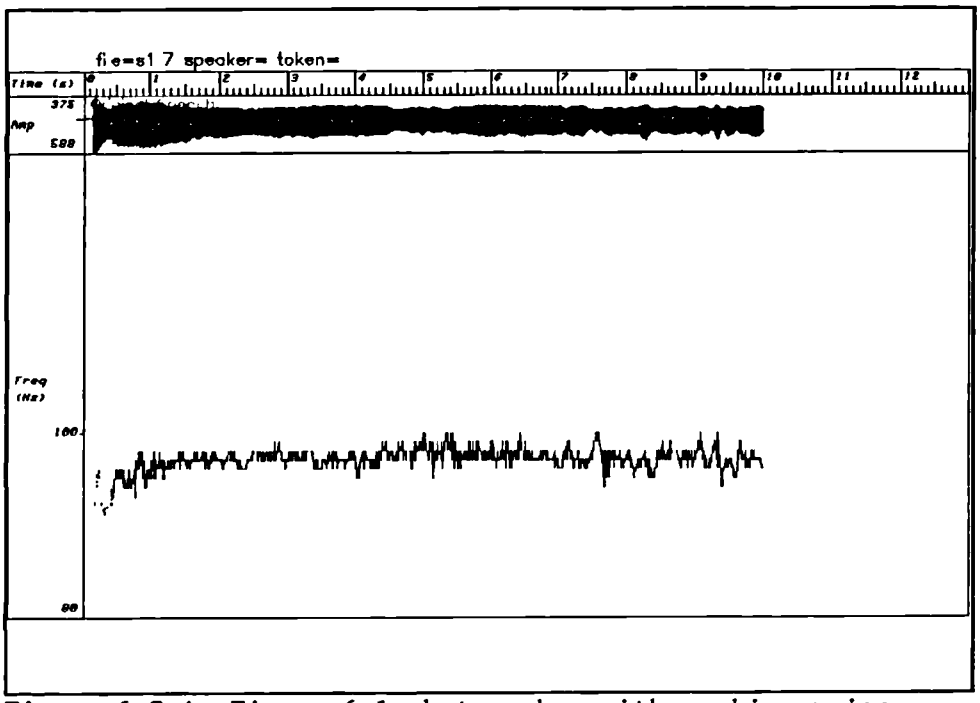


Figure 6.2 As Figure 6.1, but spoken with masking noise.

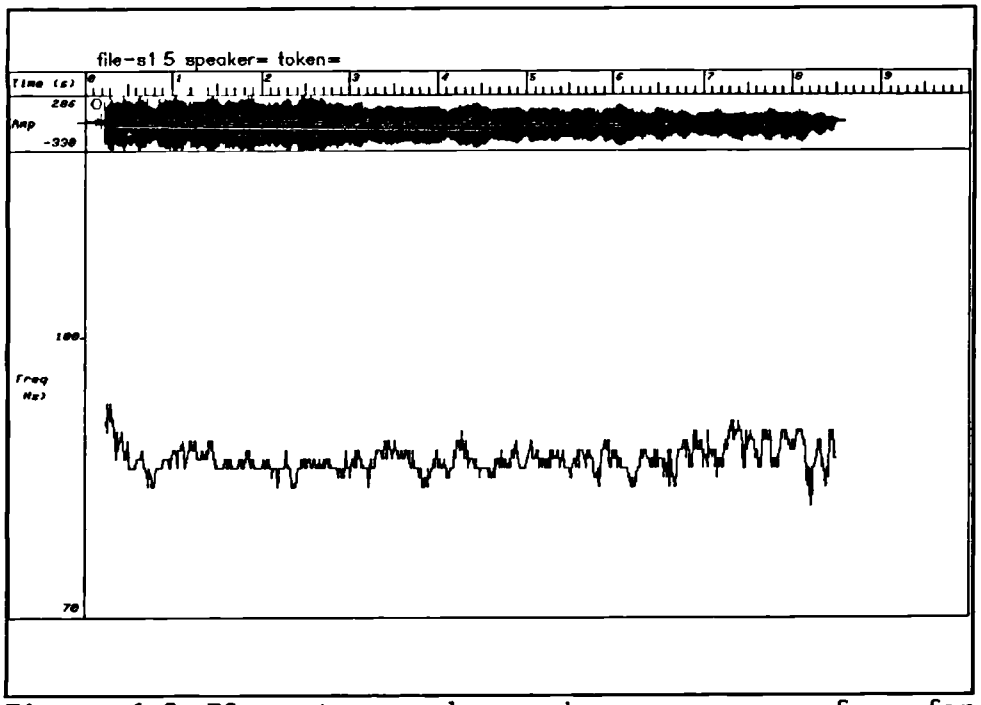


Figure 6.3 F0 contour and speech pressure waveform for sustained vowel utterance for speaker B, without masking noise. Note F0 axis: 70-120Hz.

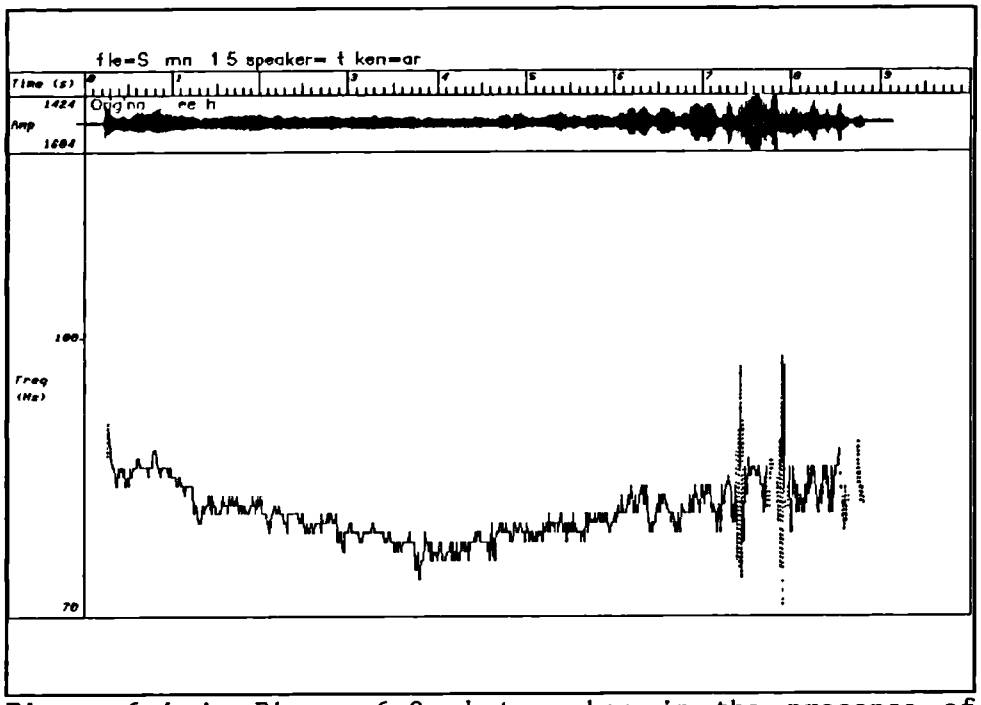


Figure 6.4 As Figure 6.3, but spoken in the presence of masking noise.

Histograms were drawn up of the distribution of Fx during the sustained vowel utterances. In addition, the F0 contours were low-pass filtered by convolution with a 500ms Hamming window, and the trends in the contours assessed across conditions.

6.2.3.2 Sentence data

Initial analysis of the sentence data was more complicated. An automatic method was devised to identify salient points for analysis in the F0 contour. These correspond closely to the salient points forming the basis for the model of pitch prominence of an accent configuration introduced in Chapter 5. This automatic method relied on an initial stylisation of the F0 contour using the 'simplified spline' algorithm of Mead (1974), against which the following turning points were labelled :

TA (Start of prepeak contextual contour, Accent 1)
TB (End of prepeak contextual contour Accent 1)
P1 (Peak of Accent 1)
T1 (Start of interaccentual trough)
T2 (End of interaccentual trough)
P2 (Peak of Accent 2)
T3 (Start of postpeak contextual contour, Accent 2)
T4 (Start of postpeak contextual contour, Accent 2)

Where the points T1 and T2 were too distant from those peaks in the stylised contour, spectral criteria were used to fix them closer: the nasal /m/ at the start of the word "may" was identified using threshold detection of the relative amplitude of a nasal formant in a synthetic transform of the speech following formant analysis using the 'rso' program of the ILS package. Similarly, the semivowels /y/ and /w/ in the words "yew" and "willow" of the short and long sentence respectively were identified using spectral analysis. Points T1 and T2 were tied to the nasal and semivowel respectively in each sentence.

In Figure 6.5, the salient points computed by the algorithm are displayed against the speech pressure waveform and the Fx contour for one instance of the short sentence spoken by Speaker A.

The acquired Fx contour was not always complete, possibly as a result of noise on the line, possibly as a result of electrode movement. In such cases, a Nominal VP contour was instead computed according to Terhardt's algorithm discussed in Chapter 1 - this can be considered equivalent to a F0 contour. Such a case is shown in Figure 6.6, for speaker B, for an instance of the short sentence spoken in the presence of masking noise.

For speaker C, the Lx data on the audio tape were accidentally overwritten. Consequently, all the F0 data for this speaker in the noise masking conditions had to be computed by Terhardt's algorithm. At the same time, there was no need to align the speech pressure waveforms in separate files to link up the F0 data with the administered noise data. An example of the relevant data for this speaker appears in Figure 6.7, for the long utterance

spoken in the presence of selective masking noise (where it cuts out during the accents).

The requirement to do spectral analysis on the speech data meant that the bleeps emitted by the local front-end computer when a speaker transgressed their locally-determined amplitude threshold had to be excised somehow from the speech signal (though this wasn't a problem for speaker C who, being the first subject, performed the experiment without this additional constraint on his production of intonation). As a consequence, a highly simplistic bleep-remover program was written, which incorporated an empirical spectral model of the emitted bleep, which was then subtracted from the overall signal. This worked well for speaker B, but less well for speaker A. This turned out not to be a problem, since speaker A indicated after the experiment that he had been able to hear the pitch of his voice during the utterance of the sentence data, though not during the utterance of the sustained vowel, so his sentence data had largely to be ignored.

After the salient points had been identified, two sorts of analysis were performed on the sentence data. Firstly, the mean F0 values and their standard deviation on all the salient points were computed, along with the mean interaccentual trough duration and mean interaccentual slope. Secondly, graphical analysis of the course of a normalised transform of F0 during the interaccentual trough was performed, in order to reveal how the course of F0 during that trough was affected by the administration of noise, and at what points the speaker had tended to transgress their amplitude threshold in production.

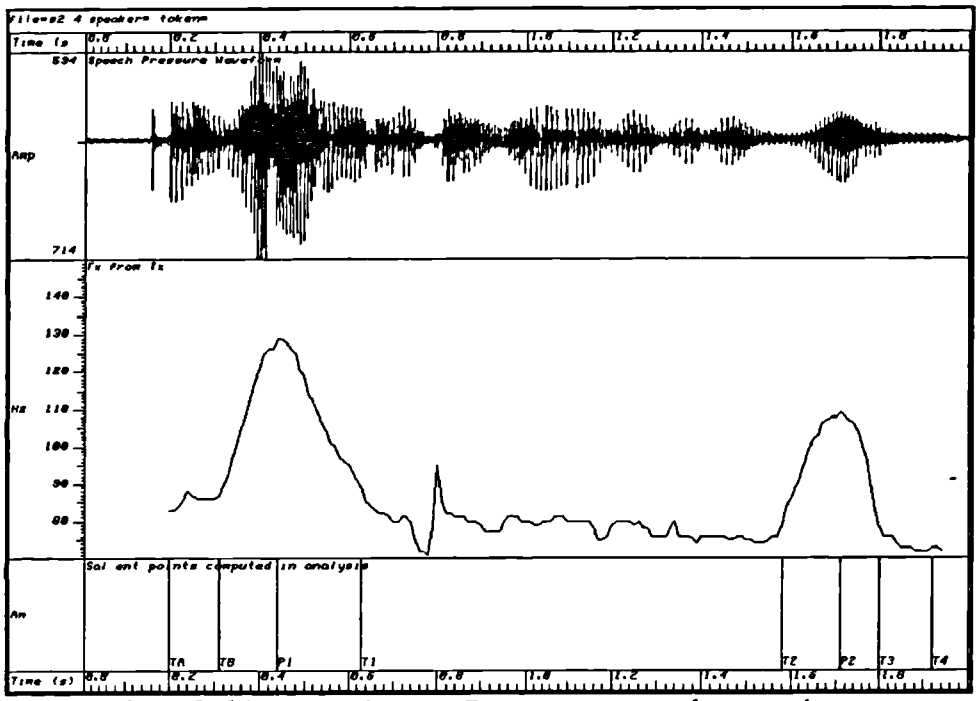


Figure 6.5 Salient points, Fx contour and speech pressure waveform in utterance of short sentence by speaker A, without noise masking.

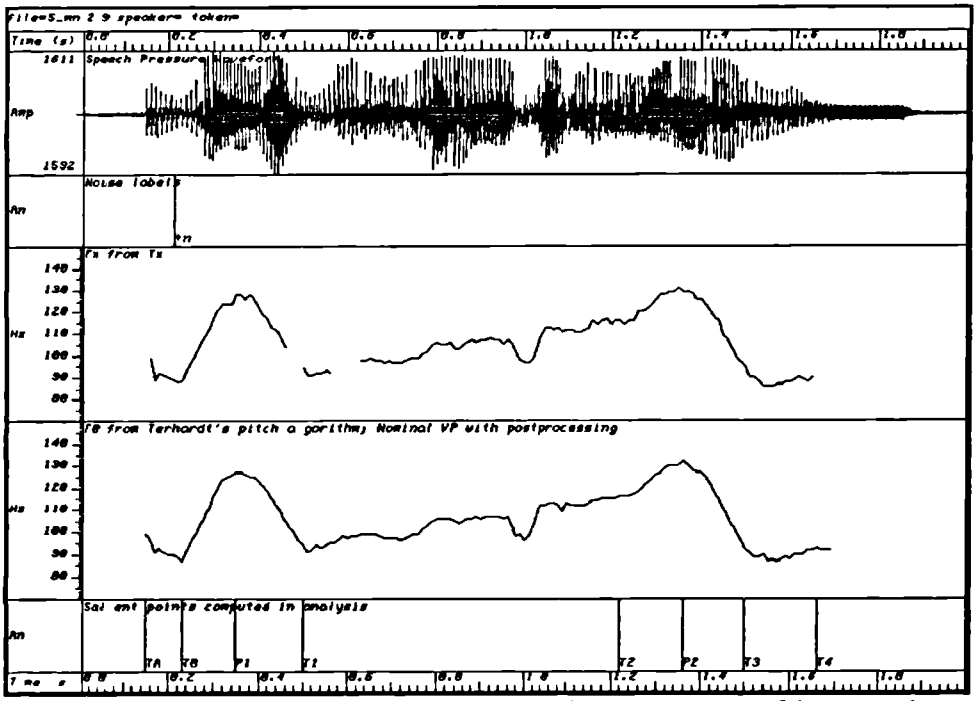


Figure 6.6 Incomplete Fx contour, F0 contour, salient points , administered noise labels and Speech pressure waveform for short sentence utterance in continuous masking noise (Speake r B).

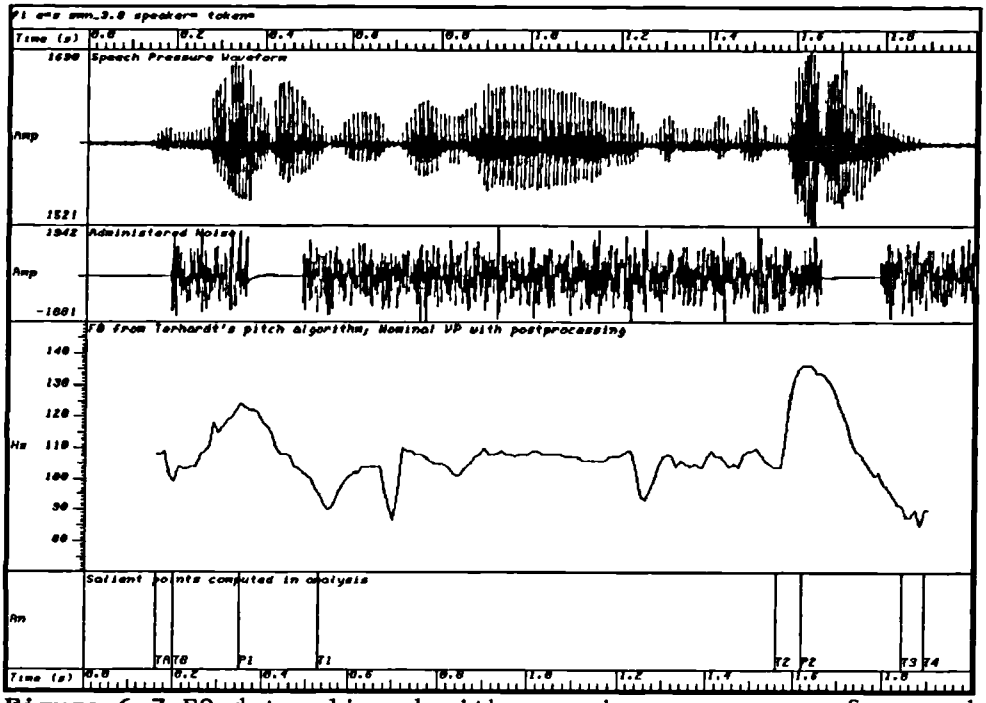


Figure 6.7 F0 data aligned with speech pressure waveform and administered noise signal during selective noise masking of utterance of long sentence by Speaker C.

6.2.3 Results and Discussion

6.2.3.1 Sustained Vowels

(It should be noted that stretches of Fx which contained creak or transient segmental coarticulation effects were excised prior to analysis. The dotted parts of the contour in Figure 6.4, for instance, represent such excisions.)

The Fx distribution analyses for the three speakers' utterances of the sustained vowels appear in Figures 6.8-6.13. At first sight it appears that no conclusive picture of the effect of noise across speakers emerges from these histograms. For two of the speakers (B and C), there is a wider spread in the distribution during continuous noise masking, but for the other, there is a narrower spread. However, the distribution of speaker A in the absence of masking noise is close to being trimodal, and it was noticeable that the speaker's choice of fundamental frequency at which to sustain phonation varied throughout the test without masking noise, starting low and ending high. Thus a direct comparison of the shapes of the distributions is not helpful in his case (although it might be thought that each 'constituent'

distribution in the trimodal distribution (Figure 6.8) is narrower than that during noise masking (Figure 6.9)).

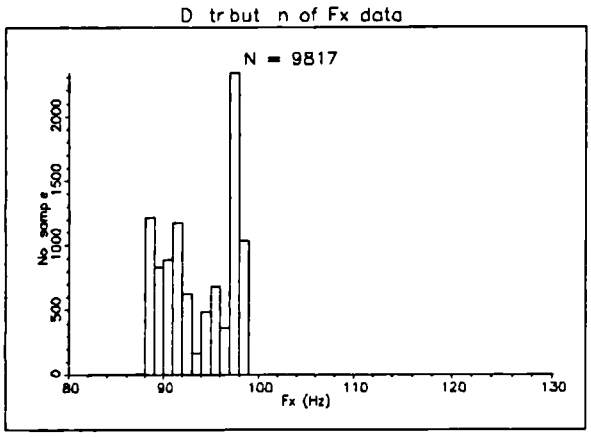


Figure 6.8 Distribution of Fx data in sustained vowel utterances without masking noise. Speaker A.

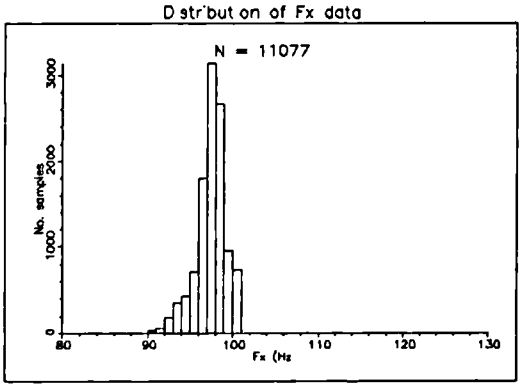


Figure 6.9 Distribution of Fx data in sustained vowel utterances with masking noise. Speaker A.

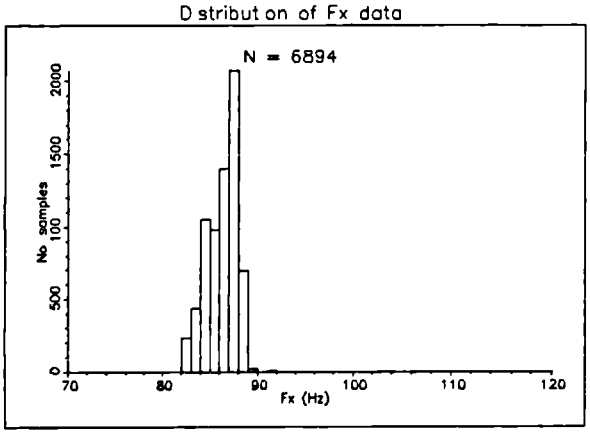


Figure 6.10 Distribution of Fx data during sustained vowel utterances, without masking noise. Speaker B.

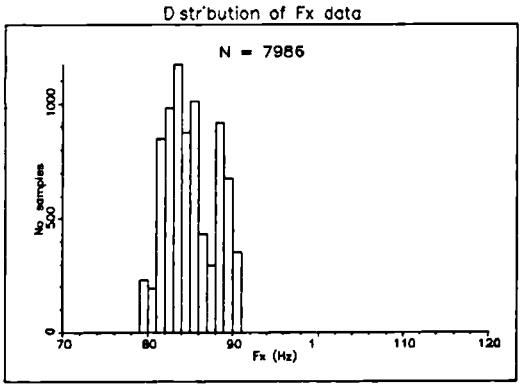


Figure 6.11 Distribution of Fx in sustained vowel utterances, with masking noise. Speaker B.

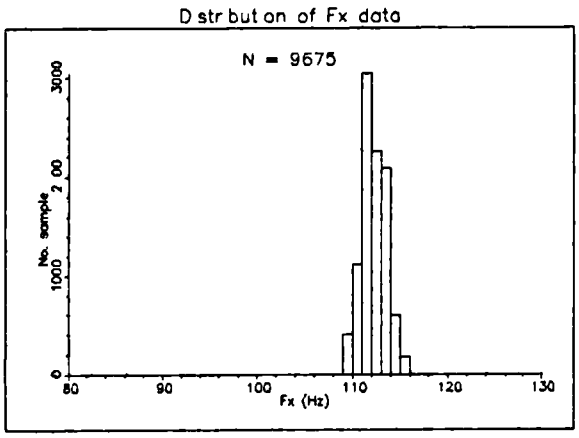


Figure 5.12 Distribution of Fx data in sustained vowel utterances, no masking noise. Speaker C.

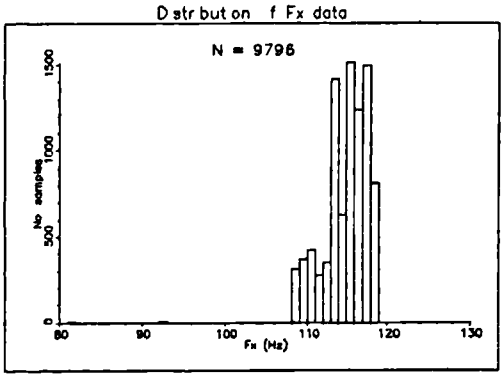


Figure 6.13 Distribution of Fx data in sustained vowel utterances, with masking noise. Speaker C.

The generally wider distribution of Fx data during noise feedback shows that it is likely that some effect is occurring such that speakers are either less able to maintain a constant pitch than with the aid of auditory feedback, or are more prone to varying the pitch intentionally, possibly, with the purpose of combatting the masking effect of the noise.

In order to get a clearer picture of the behaviour of the subjects' Fx during the two different conditions, the following analysis was performed:

The Fx data was low-pass filtered by convolution with a 500ms Hamming window. The resulting smoothed contours were analysed for local monotonic trends. The relative aggregate durations of falling and rising such trends was then determined for each of the two noise administration conditions , for each subject. The results were as follows:

Speaker	%ge dur. of speech which was falling, with no noise masking	%ge dur. of speech which was falling, with noise masking
A	59%	53.4%
B	39.6%	64.8%
C	32.5%	60%

If any of the subjects were showing more variation in their Fx data during administered masking noise as a result of an attempt to combat the effect, then one might expect that 50% of the local trends would be falling and 50% rising (wholly level local trends, which could only occur in the case of the whole filtered contour being completely level, were apportioned equally to falling and rising factors in the analysis). This result is approached by only one the speakers, Speaker A, though it should be rememembered that the slope of the trend-lines is not a factor in the analysis. It could be argued that this is evidence for disruption of the basic feedback mechanism which maintains the slope of the change in the rate of vocal fold vibration above the trivial value of zero in the case of the activity of sustained phonation.

That is possible, but there is more to be accounted for in the data. Figures 6.14 to 6.19 show the smoothed contours for each of the administered noise conditions for all three speakers. The most obvious feature of these data is that none of the speakers demonstrates a consistent overall trend in each of the conditions. Sometimes, a monotonic falling or rising trend is displayed, but equally often, complex nonmonotonic trends are displayed, both with and without continuous masking noise (although, it must be said, the amplitude of the trend movements tends to be greater in the masking noise condition, as is reflected in the histogram data above). The most characteristic feature of the trends in the masking noise data, apart from the predominance of falling trends in the case of two speakers, is the comparative magnitude of onset and offset movements between the masked and unmasked data. The larger such movements in the case of the masked stimuli possibly reflect (a) in the case of the utterance-initial movements, periods of adjustment to the sudden removal of the capacity for auditory feedback, although adjustment to what is not entirely clear, since the initial local trends can be both falling and rising, and (b) in the case of the utterance-final movements, either premature relinquishment to passive mechanisms determining the course of Fx, for falling movements, or excessive use of the expiratory musculature to maintain phonation towards the end of the utterance, resulting in heightened transglottal pressure, for rising movements.

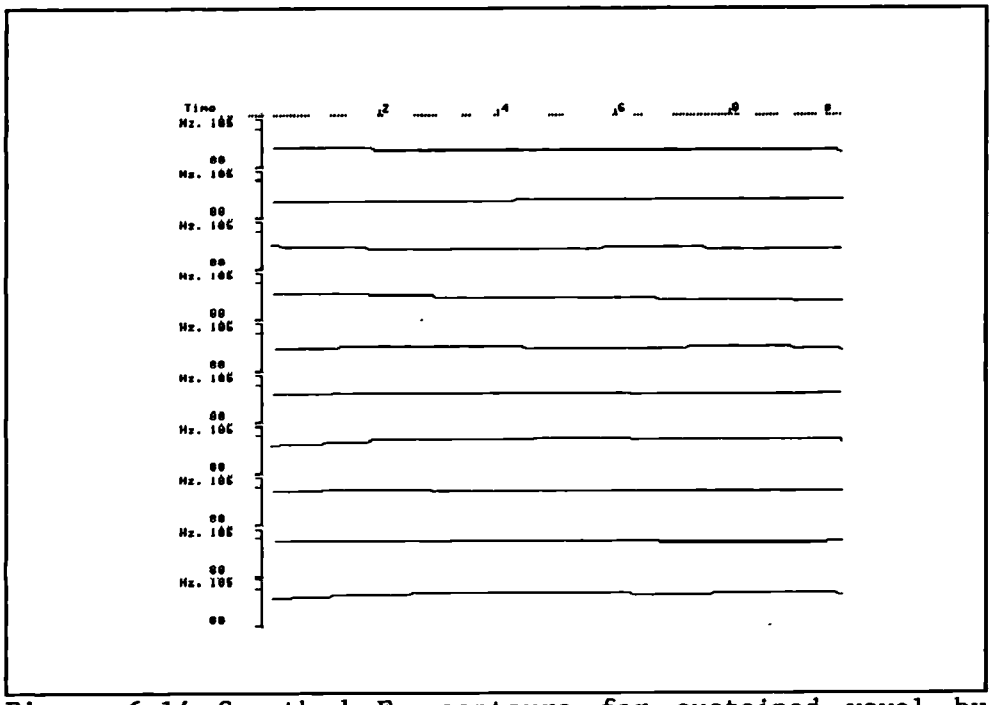


Figure 6.14 Smoothed Fx contours for sustained vowel by SPEAKER A, without masking noise.

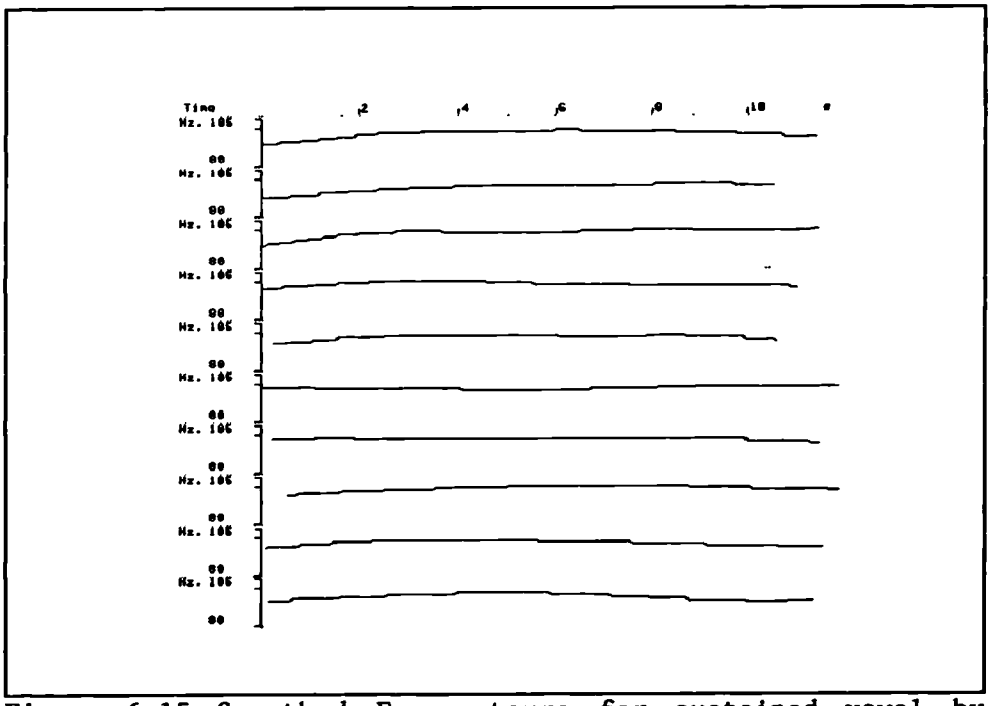


Figure 6.15 Smoothed Fx contours for sustained vowel by SPEAKER A, with masking noise.

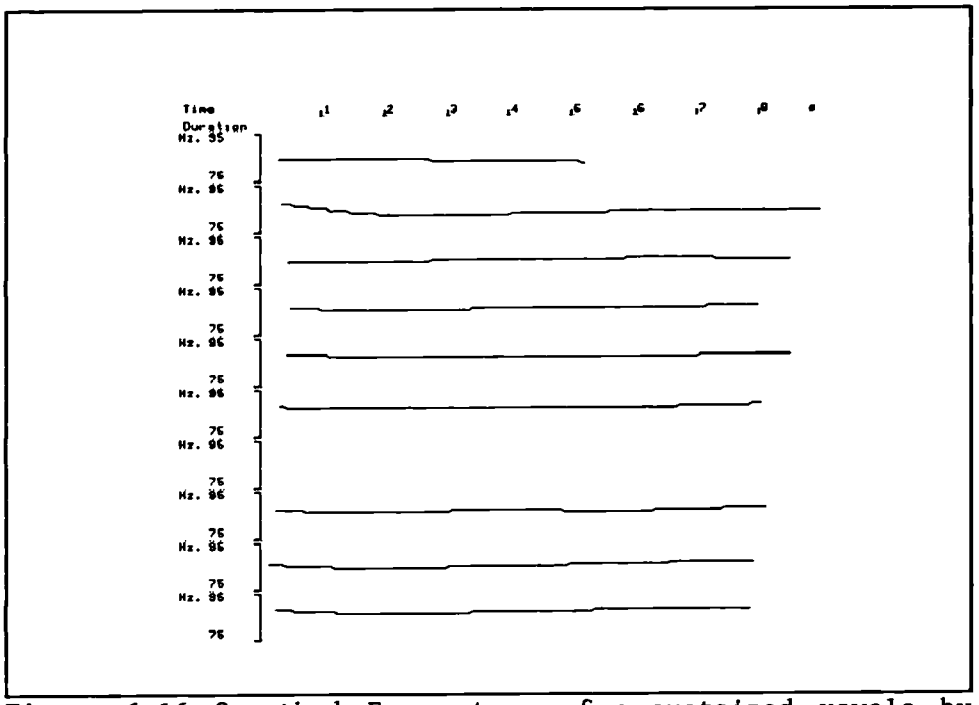


Figure 6.16 Smoothed Fx contours for sustained vowels by SPEAKER B, without masking noise. The seventh contour was lo st through experimental error.

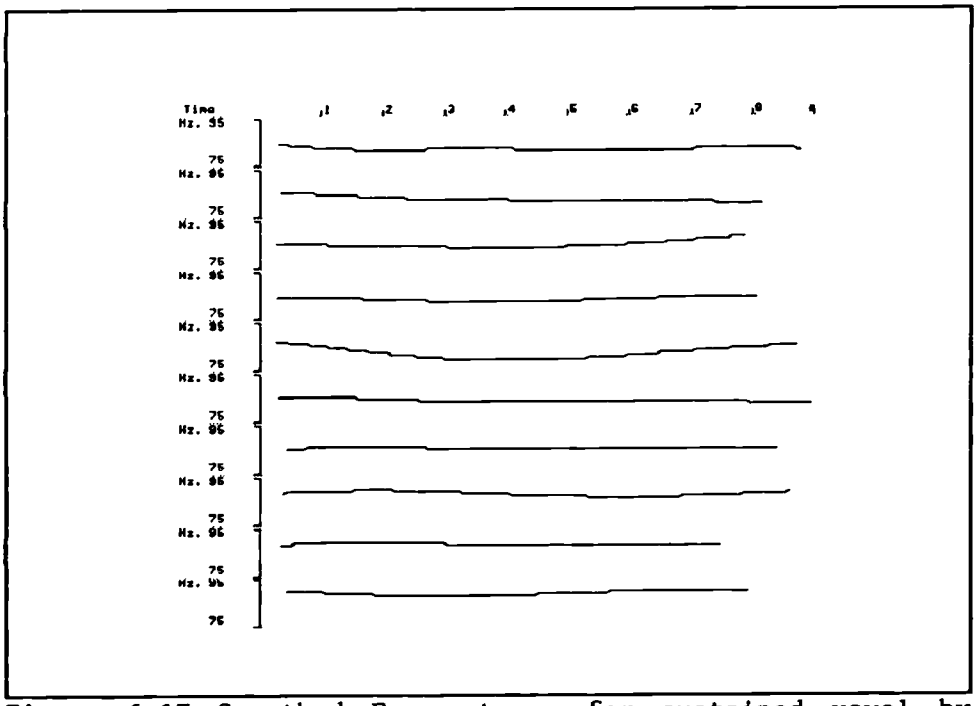


Figure 6.17 Smoothed Fx contours for sustained vowel by SPEAKER B, with masking noise.

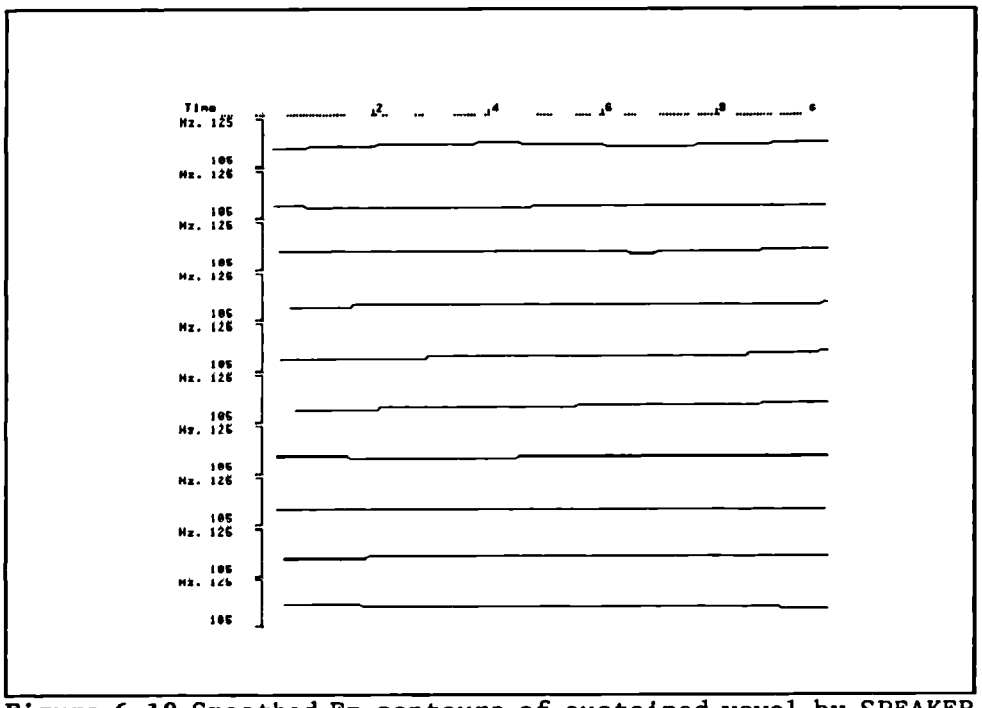


Figure 6.18 Smoothed Fx contours of sustained vowel by SPEAKER C, without masking noise.

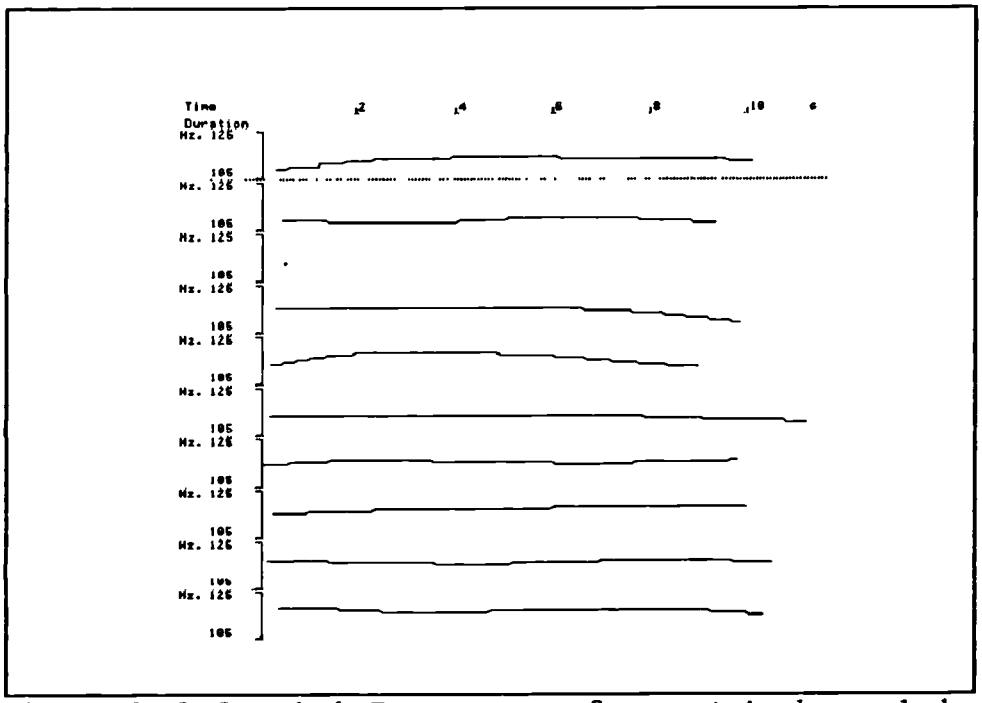


Figure 6.19 Smoothed Fx contours for sustained vowel by SPEAKER C, with masking noise. The third contour was lost through experimental error.

6.2.3.2 Sentence Data

The analysis of the effects of the different sorts of administered masking noise on the course of Fx during the production of the sentence data revealed that there is one factor which predominated in the course of F0 during any of the forms of masking noise, and this was the tendency for speakers to show a gradual increase in the course of Fx as the length of adminstration of masking noise increased. This was true of all speakers, though less of some than others. For instance, Speaker A shows little difference in the course of his interaccentual F0 contour in the normalised plots appearing in Figures 6.20 and 6.21. But speaker B shows a marked difference in the interaccentual stretch in the presence of continuous masking noise - the trend in the data is strongly rising in Fig. 6.23 as compared with Fig 6.22. And even Speaker A (for whom the masking noise was inadequate to mask completely his perception of the pitch of his own voice) reveals a change in the difference between the F0 value of first peak and F0 value of second peak from 20.6 and 20.5 Hz respectively for short and long sentences in the absence of masking noise, to 2.5 and -5Hz for the same in the presence of continuous masking noise (see data analysis in section 6.4.1).

To summarize the relevant data for the three speakers:

Speaker	Peak difference	Trough difference	Masking Condition	Sentence
A	-20.6Hz	-9.2Hz	No noise	S
A	-20.5Hz	-12.2Hz	No noise	L
A	-2.5Hz	-9.4Hz	Cont.noise	S
A	-4.4Hz	+4.3Hz	Cont.noise	L
B	-7.0Hz	+8.7Hz	No noise	S
B	-1.6Hz	+11.6Hz	No noise	L
B	+5.0Hz	+16Hz	Cont.Noise	S
B	+4.8Hz	+23.6Hz	Cont.Noise	L
B	-2.11Hz	+13.2Hz	Sel. Noise	S
B	-	-	Sel.Noise	L
B	-5.2Hz	+13.4Hz	Sel.Noise(F)	S
B	0.0Hz	+12.7Hz	Sel.Noise(F)	L

Speaker	Peak difference	Trough difference	Masking Condition	Sentence
C	+3.8Hz	-2.6Hz	No noise	S
C	+0.5Hz	-2.1Hz	No noise	L
C	+4.3Hz	-1.9Hz	Cont.noise	S
C	+8.0Hz	+4.3Hz	Cont.noise	L
C	+5.2Hz	-0.4Hz	Sel.Noise	S
C	+6.9Hz	+3.0Hz	Sel.Noise	L
C	+7.6Hz	+6.8Hz	Sel.Noise(F)	S
C	-	-	Sel.Noise(F)	L

(where Peak Difference = P2-P1 and Trough Difference = T2-T1)
For all three speakers, the values of the later points in the configuration, be it accent or context, increase with the administration of masking noise. Thus, what produces an interesting and apparently complex set of laryngeal behaviour in the production of sustained vowels produces a pretty unary sort of behaviour in the production of sentence data.

It seems pretty certain that the reason for this is that one of the concomitants of increase in speech amplitude is an increase in F0. The former comes about in the presence of masking noise because of the Lombard Effect - so-named after the Frenchman who discovered and researched it. This is taken to be an almost reflex response to being made to speak in noisy surroundings. However, Lane and Tranel (1971) suggest that the effect is attenuated in the absence of a context of communication. In fact, there is a slight increase in level noticeable in some of the sustained vowel data. It is perhaps not so strong as in the case of the sentence data for the very reason that Lane and Tranel suggest.

6.3 Conclusion

The aim of this experiment was to make some exploratory investigations into the form of intonation contours, notably unaccented stretches of contour, in the absence of auditory feedback. The experimental protocol used was seen to be appropriate in the case of the production of sustained vowels, in that further questions were stimulated about the process of speech production in respect of the maintenance of a particular rate and, a fortiori, change in rate of vocal fold vibration. However, the maintenance of a constant pitch on a

sustained vowel is not characteristically an intonation task. When the protocol was used in the case of tasks which do require the use of intonation, the results are vitiated by the existence of an automatic mechanism which increases the level of a speaker's voice in the presence of noise, and with it, the pitch of that voice. In the following chapter, one or two suggestions are made as to how the experimental protocol could be improved in later research in order to counteract that vitiating factor.

6.3 APPENDIX 1

Instructions to subjects for the production exeriment in this chapter:

PRODUCTION EXPERIMENT

In this experiment, you will be asked to utter three things a number of times under varying conditions: a sustained "ar" vowel, and two sentences. The conditions are

(i) normal

(ii) with masking noise throughout the utterance.

(iii) with selective masking noise during the utterance.

(iv) as (iii), but with speech feedback during those times when there is no masking noise.

In all conditions you should speak wearing headphones and Laryngograph electrodes. Your speech will be recorded.

(i) Normal

In this condition, you utter the following, each ten times:

1. A sustained /ar/ vowel, at constant pitch, for 10 seconds or as long as you can manage.
2. The sentence 'A willow may be rarer than a yew'.
   You should imagine the context to be something like the following: 'On riverbanks, you can often find many of the best-loved English trees; willows, aspens and ashes. But then think of the forest. There, **a willow may be rarer than a yew.**'
   You should accentuate only the words 'willow' and 'yew'.
3. The sentence 'A mallow may be rarer even than a willow'. You could imagine this as a rider to the previous passage:'...and in the east of the country, **a mallow may be rarer even than a willow** '. Similarly, you should accentuate only the words 'mallow' and 'willow'.

After you have spoken these sentences, there will be a short period of analysis, in order to determine parameters for the next three conditions.

After the analysis, you will be asked to sat the sustained vowel and the two test sentences in the presence of masking noise, but maintaining the level at which you spoke them in the normal condition. To aid you in this task, an intensity meter indicating the level above which you shouldn't go will be displayed on the computer screen. If you exceed this level, a bar will flash inside the no-go area, and you'll hear a beep. The level of masking noise will be adjusted until you cannot detect the pitch of your voice.

The degree of noise being administered at this level will then be determined using an artificial ear feeding a spectrum analyzer.

Next, you should utter the same things (again ten times each) under the three different conditions of noise masking:

(ii) with masking noise throughout the utterance
In this condition, when you speak, your speech will be continuously masked by noise. You should try to maintain your speech at a level below that indicated by the intensity meter on the screen.

(iii) with selective masking noise during the utterance
In this condition, your speech will be masked only during the unaccentuated stretches. You should again maintain your speech below the level indicated.

(iv) with selective masking noise during the utterance, and with speech feedback
In this condition, your speech will be masked during the unaccentuated stretches, and at other times your speech will be fed back to you over the headphones. Again, you should maintain the level of your speech below that indicated.

Conduct of the experiment is via a menu interface on the computer screen, which will be explained to you.

Thank you very much for your cooperation.

6.4 APPENDIX 2. Summary results of the sentence production part of the production experiment.

In this section, two sets of data are presented
(1) the mean Fx values (or F0 values when Terhardt's pitch algorithm was used, notably in the case of Subject C) and their standard deviations are presented of the salient points automatically determined in the analysis, and in addition the difference in frequency between P1 and P2 (P1-P2), the duration between P1 and P2 (P1->P2), the slope determined by linear regression between T1 and T2 (T1\T2) and the slope determined in the same way between T3 and T4 (T3\T4). The conditions are presented in capitals at the start of each batch of results. The data for Speaker A was only partially determined after his observation that the noise masking was insufficiently effective for him. Two batches of data were unavailable at the time of writing, one for speaker B and one for speaker C. These are marked UNAVAILABLE beneath the respective heading.

(2) Some example plots demonstrating graphical display of mean values of a normalised transform of the Fx/F0 data during the interaccentual trough (T1->T2) for particular conditions. The numbers in the plots represent the proportion of analysed tokens which contribute to the value at that point from 0 (<= 10%) to 9 (100%). The height of any solid vertical bars represents the number of tokens which were accompanied by masking noise at a particular point in the averaged interaccentual contour, and the plus-signs ('+') indicate episodes of overload which resulted in the subject being beeped by the computer.

6.4.1 SUMMARY ANALYSIS OF FO VALUES ON SALIENT POINTS
SHORT SENTENCES - NO MASKING NOISE - SPEAKER A
Summary of statistics over 10 tokens

	Mean	Standard Deviation
TA	83.00 Hz	(+/-)9.661
TB	83.00 Hz	(+/-)8.679
P1	133.30 Hz	(+/-)9.661
T1	90.10 Hz	(+/-)8.679
T2	80.90 Hz	(+/-)7.424
P2	112.70 Hz	(+/-)4.175
T3	78.60 Hz	(+/-)4.864
T4	72.80 Hz	(+/-)4.029
P1-P2	20.60 Hz	(+/-)6.620
P1->P2	1.272 secs	(+/-)0.060
T1\T2	-5.635 Hz/sec	(+/-)3.105
T3\T4	-31.081 Hz/sec	(+/-)9.803

LONG SENTENCES - NO NOISE MASKING - SPEAKER A
Summary of statistics over 10 tokens

	Mean	Standard Deviation
TA	80.60 Hz	(+/-)7.589
TB	80.30 Hz	(+/-)3.234
P1	129.70 Hz	(+/-)7.589
T1	92.10 Hz	(+/-)3.234
T2	80.90 Hz	(+/-)8.757
P2	109.20 Hz	(+/-)3.573
T3	67.10 Hz	(+/-)4.254
T4	73.40 Hz	(+/-)5.432
P1-P2	20.50 Hz	(+/-)4.696
P1->P2	1.588 secs	(+/-)0.043
T1\T2	-5.162 Hz/sec	(+/-)2.294
T3\T4	-39.352 Hz/sec	(+/-)148.235

SHORT SENTENCES - CONTINUOUS NOISE FEEDBACK - SPEAKER A
Summary of statistics over 10 tokens

	Mean	Standard Deviation
TA	73.70 Hz	(+/-)12.184
TB	80.60 Hz	(+/-)3.596
P1	119.60 Hz	(+/-)12.184
T1	85.20 Hz	(+/-)3.596
T2	75.80 Hz	(+/-)6.363
P2	117.10 Hz	(+/-)4.780
T3	81.80 Hz	(+/-)5.266
T4	72.40 Hz	(+/-)5.043
P1-P2	2.50 Hz	(+/-)7.122
P1->P2	1.334 secs	(+/-)0.056
T1\T2	-6.747 Hz/sec	(+/-)2.314
T3\T4	-82.128 Hz/sec	(+/-)150.092

LONG SENTENCES - CONTINUOUS MASKING NOISE - SPEAKER A
Summary of statistics over 10 tokens

	Mean	Standard Deviation
TA	81.30 Hz	(+/-)12.807
TB	82.50 Hz	(+/-)5.255
P1	123.50 Hz	(+/-)12.807
T1	80.70 Hz	(+/-)5.255
T2	85.00 Hz	(+/-)3.808
P2	119.10 Hz	(+/-)28.783
T3	78.70 Hz	(+/-)4.243
T4	73.50 Hz	(+/-)3.213
P1-P2	4.40 Hz	(+/-)3.239
P1->P2	1.592 secs	(+/-)0.041
T1\T2	-2.499 Hz/sec	(+/-)2.268
T3\T4	-62.510 Hz/sec	(+/-)156.412

SHORT SENTENCES - NO NOISE FEEDBACK - SPEAKER B
Summary of statistics over 10 tokens

	Mean	Standard Deviation
TA	96.90 Hz	(+/-)6.951
TB	85.30 Hz	(+/-)3.020
P1	113.40 Hz	(+/-)6.951
T1	88.20 Hz	(+/-)3.020
T2	96.90 Hz	(+/-)7.321
P2	106.40 Hz	(+/-)2.616
T3	83.70 Hz	(+/-)4.122
T4	80.50 Hz	(+/-)6.720
P1-P2	7.00 Hz	(+/-)3.055
P1->P2	1.057 secs	(+/-)0.035
T1\T2	11.647 Hz/sec	(+/-)3.284
T3\T4	-38.180 Hz/sec	(+/-)21.677

LONG SENTENCES - NO NOISE FEEDBACK - SPEAKER B
Summary of statistics over 10 tokens

	Mean	Standard Deviation
TA	93.30 Hz	(+/-)8.152
TB	100.00 Hz	(+/-)8.219
P1	120.20 Hz	(+/-)8.152
T1	92.70 Hz	(+/-)8.219
T2	104.30 Hz	(+/-)4.756
P2	118.60 Hz	(+/-)2.163
T3	84.90 Hz	(+/-)3.529
T4	83.40 Hz	(+/-)6.077
P1-P2	1.60 Hz	(+/-)3.098
P1->P2	1.307 secs	(+/-)0.049
T1\T2	10.625 Hz/sec	(+/-)4.368
T3\T4	-1.343 Hz/sec	(+/-)17.127

SHORT SENTENCES - CONTINUOUS MASKING NOISE - SPEAKER B
Summary of statistics over 10 tokens

	Mean	Standard Deviation
TA	96.80 Hz	(+/-)4.211
TB	90.80 Hz	(+/-)2.573
P1	121.80 Hz	(+/-)4.211
T1	94.30 Hz	(+/-)2.573
T2	110.30 Hz	(+/-)3.676
P2	126.80 Hz	(+/-)2.497
T3	91.60 Hz	(+/-)2.830
T4	88.80 Hz	(+/-)3.765
P1-P2	-5.00 Hz	(+/-)2.789
P1->P2	1.039 secs	(+/-)0.053
T1\T2	23.745 Hz/sec	(+/-)4.352
T3\T4	-4.741 Hz/sec	(+/-)29.160

LONG SENTENCES - CONTINUOUS MASKING NOISE - SPEAKER B
Summary of statistics over 10 tokens

	Mean	Standard Deviation
TA	92.10 Hz	(+/-)11.396
TB	94.20 Hz	(+/-)2.860
P1	122.90 Hz	(+/-)11.396
T1	93.70 Hz	(+/-)2.860
T2	117.30 Hz	(+/-)4.175
P2	127.70 Hz	(+/-)3.622
T3	86.60 Hz	(+/-)4.785
T4	86.70 Hz	(+/-)11.066
P1-P2	-4.80 Hz	(+/-)11.868
P1->P2	1.204 secs	(+/-)0.039
T1\T2	27.506 Hz/sec	(+/-)5.003
T3\T4	10.174 Hz/sec	(+/-)21.301

SHORT SENTENCES - SELECTIVE MASKING NOISE - SPEAKER B
Summary of statistics over 9 tokens

	Mean	Standard Deviation
TA	97.33 Hz	(+/-)4.093
TB	89.89 Hz	(+/-)1.965
P1	120.67 Hz	(+/-)4.093
T1	91.11 Hz	(+/-)1.965
T2	104.33 Hz	(+/-)5.385
P2	118.56 Hz	(+/-)3.408
T3	88.00 Hz	(+/-)2.179
T4	85.33 Hz	(+/-)3.283
P1-P2	2.11 Hz	(+/-)5.110
P1->P2	1.088 secs	(+/-)0.054
T1\T2	20.461 Hz/sec	(+/-)4.498
T3\T4	-20.210 Hz/sec	(+/-)39.886

LONG SENTENCES - SELECTIVE MASKING NOISE - SPEAKER B

UNAVAILABLE

SHORT SENTENCES - SELECTIVE MASKING NOISE ALTERNATING WITH SPEECH FEEDBACK - SPEAKER B

Summary of statistics over 10 tokens

	Mean	Standard Deviation
TA	101.40 Hz	(+/-)5.337
TB	91.20 Hz	(+/-)5.138
P1	130.90 Hz	(+/-)5.337
T1	93.40 Hz	(+/-)5.138
T2	106.80 Hz	(+/-)5.174
P2	125.70 Hz	(+/-)4.575
T3	89.20 Hz	(+/-)5.266
T4	85.10 Hz	(+/-)5.658
P1-P2	5.20 Hz	(+/-)5.594
P1->P2	1.239 secs	(+/-)0.037
T1\T2	21.399 Hz/sec	(+/-)6.510
T3\T4	-22.021 Hz/sec	(+/-)44.927

LONG SENTENCES - SELECTIVE MASKING NOISE ALTERNATING WITH SPEECH FEEDBACK - SPEAKER B

Summary of statistics over 9 tokens

	Mean	Standard Deviation
TA	108.67 Hz	(+/-)7.331
TB	101.89 Hz	(+/-)3.887
P1	129.44 Hz	(+/-)7.331
T1	96.00 Hz	(+/-)3.887
T2	108.67 Hz	(+/-)5.053
P2	129.44 Hz	(+/-)4.301
T3	87.44 Hz	(+/-)2.500
T4	86.33 Hz	(+/-)4.927
P1-P2	0.00 Hz	(+/-)7.984
P1->P2	1.350 secs	(+/-)0.068
T1\T2	16.116 Hz/sec	(+/-)1.735
T3\T4	3.644 Hz/sec	(+/-)15.250

SHORT SENTENCES - NO MASKING NOISE - SPEAKER C

Summary of statistics over 10 tokens

	Mean	Standard Deviation
TA	104.10 Hz	(+/-)6.367
TB	88.10 Hz	(+/-)2.079
P1	124.60 Hz	(+/-)6.367
T1	99.60 Hz	(+/-)2.079
T2	97.00 Hz	(+/-)5.777
P2	128.40 Hz	(+/-)2.633
T3	87.20 Hz	(+/-)3.712
T4	81.80 Hz	(+/-)5.420
P1-P2	-3.80 Hz	(+/-)9.864
P1->P2	1.201 secs	(+/-)0.102
T1\T2	-1.718 Hz/sec	(+/-)3.030
T3\T4	-79.153 Hz/sec	(+/-)72.838

LONG SENTENCES - NO MASKING NOISE - SPEAKER C
Summary of statistics over 10 tokens

	Mean	Standard Deviation
TA	102.80 Hz	(+/-)4.662
TB	97.40 Hz	(+/-)2.503
P1	126.00 Hz	(+/-)4.662
T1	98.00 Hz	(+/-)2.503
T2	100.10 Hz	(+/-)3.916
P2	126.50 Hz	(+/-)4.320
T3	88.00 Hz	(+/-)3.213
T4	85.20 Hz	(+/-)3.749
P1-P2	-0.50 Hz	(+/-)3.567
P1->P2	1.350 secs	(+/-)0.055
T1\T2	6.411 Hz/sec	(+/-)2.360
T3\T4	-24.985 Hz/sec	(+/-)39.022

SHORT SENTENCES - CONTINUOUS MASKING NOISE - SPEAKER C
Summary of statistics over 10 tokens

	Mean	Standard Deviation
TA	119.50 Hz	(+/-)11.306
TB	90.80 Hz	(+/-)4.662
P1	119.60 Hz	(+/-)11.306
T1	101.50 Hz	(+/-)4.662
T2	99.60 Hz	(+/-)5.060
P2	123.90 Hz	(+/-)4.035
T3	93.60 Hz	(+/-)5.873
T4	86.90 Hz	(+/-)8.412
P1-P2	-4.30 Hz	(+/-)8.407
P1->P2	1.181 secs	(+/-)0.064
T1\T2	4.244 Hz/sec	(+/-)4.679
T3\T4	-129.102 Hz/sec	(+/-)117.490

LONG SENTENCES - CONTINUOUS MASKING NOISE - SPEAKER C
Summary of statistics over 10 tokens

	Mean	Standard Deviation
TA	112.00 Hz	(+/-)7.688
TB	97.30 Hz	(+/-)6.816
P1	118.50 Hz	(+/-)7.688
T1	101.50 Hz	(+/-)6.816
T2	105.80 Hz	(+/-)4.720
P2	126.50 Hz	(+/-)3.689
T3	94.90 Hz	(+/-)4.541
T4	87.10 Hz	(+/-)3.171
P1-P2	-8.00 Hz	(+/-)4.989
P1->P2	1.356 secs	(+/-)0.072
T1\T2	4.824 Hz/sec	(+/-)4.371
T3\T4	-146.554 Hz/sec	(+/-)151.468

SHORT SENTENCES - SELECTIVE MASKING NOISE - SPEAKER C
Summary of statistics over 10 tokens

	Mean	Standard Deviation
TA	120.50 Hz	(+/-)6.621
TB	95.60 Hz	(+/-)3.836
P1	125.00 Hz	(+/-)6.621
T1	99.70 Hz	(+/-)3.836
T2	99.30 Hz	(+/-)2.539
P2	130.20 Hz	(+/-)3.234
T3	96.90 Hz	(+/-)4.990
T4	89.10 Hz	(+/-)3.190
P1-P2	-5.20 Hz	(+/-)2.821
P1->P2	1.068 secs	(+/-)0.050
T1\T2	5.925 Hz/sec	(+/-)4.201
T3\T4	-131.929 Hz/sec	(+/-)132.168

LONG SENTENCES - SELECTIVE MASKING NOISE - SPEAKER C
Summary of statistics over 8 tokens

	Mean	Standard Deviation
TA	113.88 Hz	(+/-)5.793
TB	102.00 Hz	(+/-)5.451
P1	120.88 Hz	(+/-)5.793
T1	99.75 Hz	(+/-)5.451
T2	102.75 Hz	(+/-)4.549
P2	127.75 Hz	(+/-)4.559
T3	91.50 Hz	(+/-)5.418
T4	86.63 Hz	(+/-)7.686
P1-P2	-6.88 Hz	(+/-)10.960
P1->P2	1.266 secs	(+/-)0.032
T1\T2	4.360 Hz/sec	(+/-)3.167
T3\T4	-109.670 Hz/sec	(+/-)161.247

SHORT SENTENCES - SELECTIVE MASKING NOISE ALTERNATING WITH SPEECH FEEDBACK - SPEAKER C
Summary of statistics over 10 tokens

	Mean	Standard Deviation
TA	122.60 Hz	(+/-)12.492
TB	102.30 Hz	(+/-)4.373
P1	125.30 Hz	(+/-)12.492
T1	101.40 Hz	(+/-)4.373
T2	108.20 Hz	(+/-)2.497
P2	132.90 Hz	(+/-)3.596
T3	98.60 Hz	(+/-)3.736
T4	85.50 Hz	(+/-)4.175
P1-P2	-7.60 Hz	(+/-)4.502
P1->P2	1.126 secs	(+/-)0.057
T1\T2	7.789 Hz/sec	(+/-)4.728
T3\T4	-93.639 Hz/sec	(+/-)100.346

LONG SENTENCES - SELECTIVE MASKING NOISE ALTERNATING WITH SPEECH FEEDBACK - SPEAKER C

UNAVAILABLE

6.4.2 EXAMPLE AVERAGED INTERACCENTUAL CONTOUR PLOTS

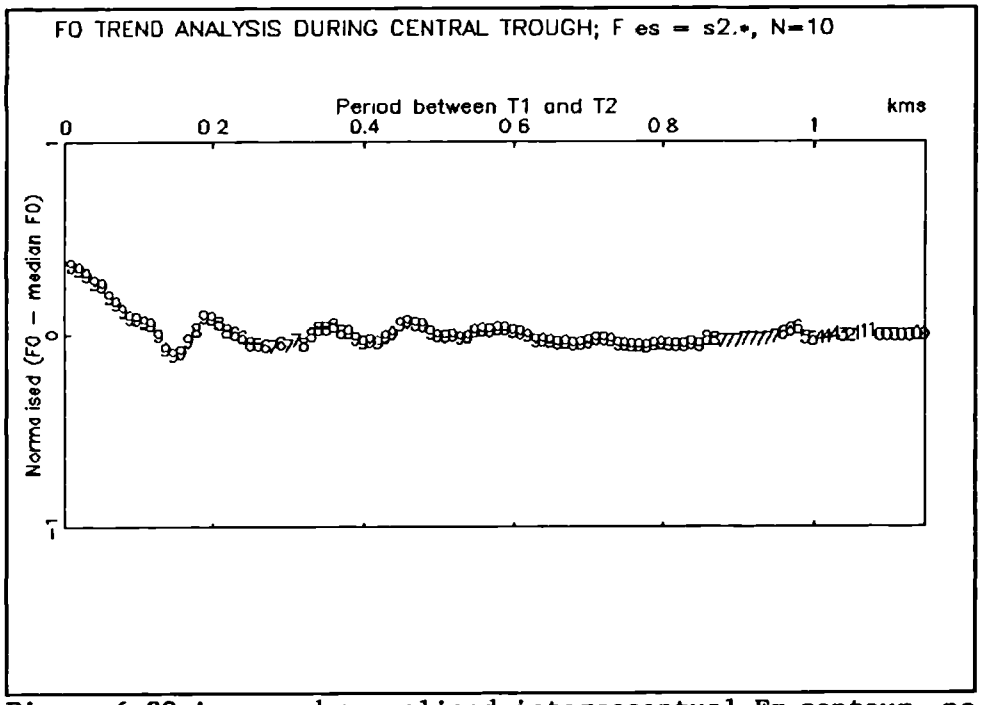


Figure 6.20 Averaged normalised interaccentual Fx contour, no masking noise, SPEAKER A. Short sentence data.

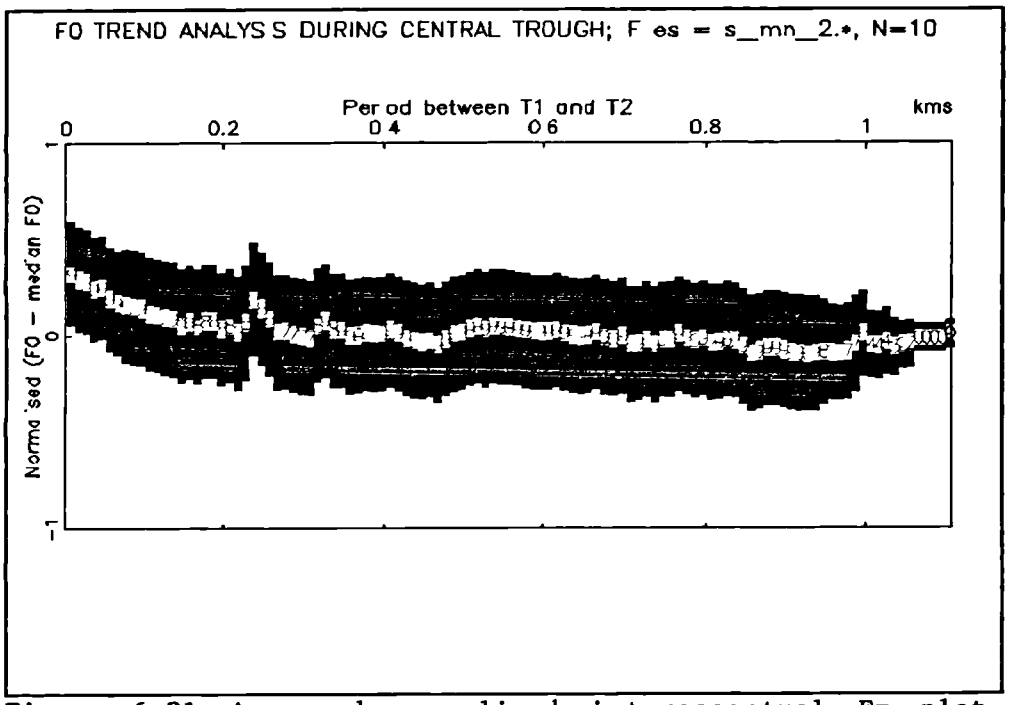


Figure 6.21 Averaged normalised interaccentual Fx plot, continuous masking noise; SPEAKER A.

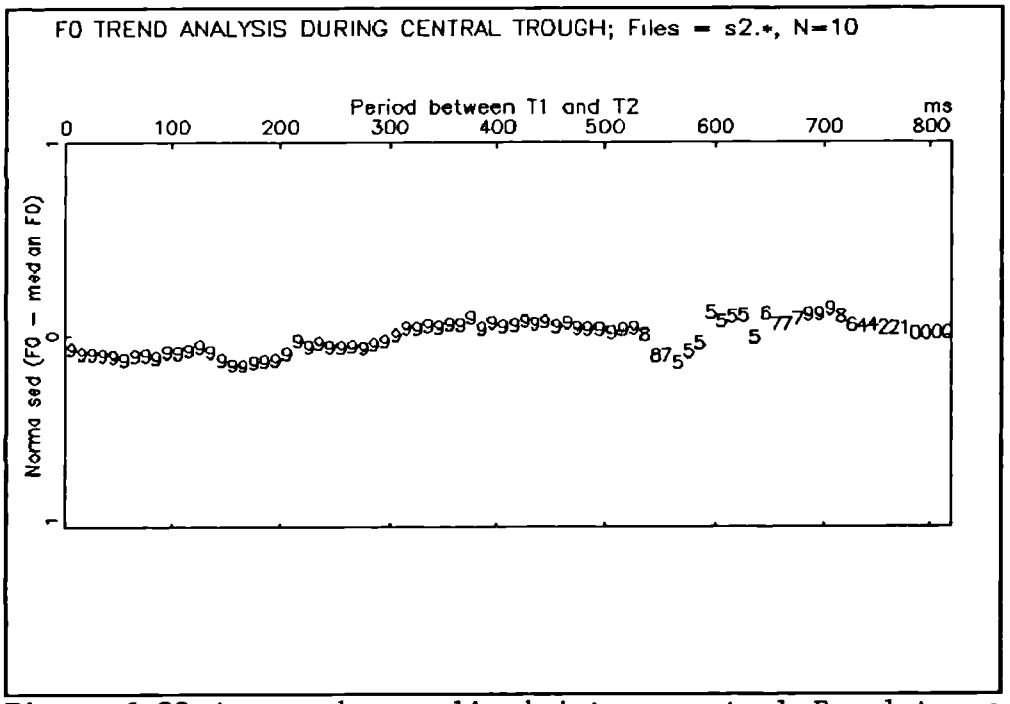


Figure 6.22 Averaged normalised interaccentual Fx plot, no masking noise, SPEAKER B. Short sentence data.

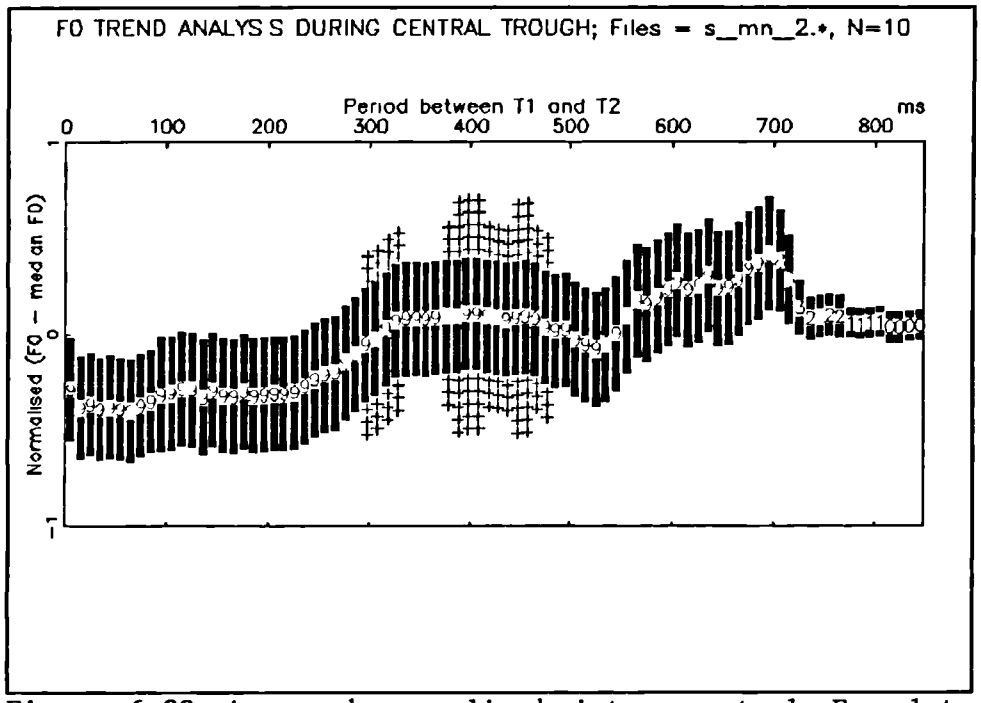


Figure 6.23 Averaged normalised interaccentual Fx plot, continuous masking noise, SPEAKER B. Short sentence data.

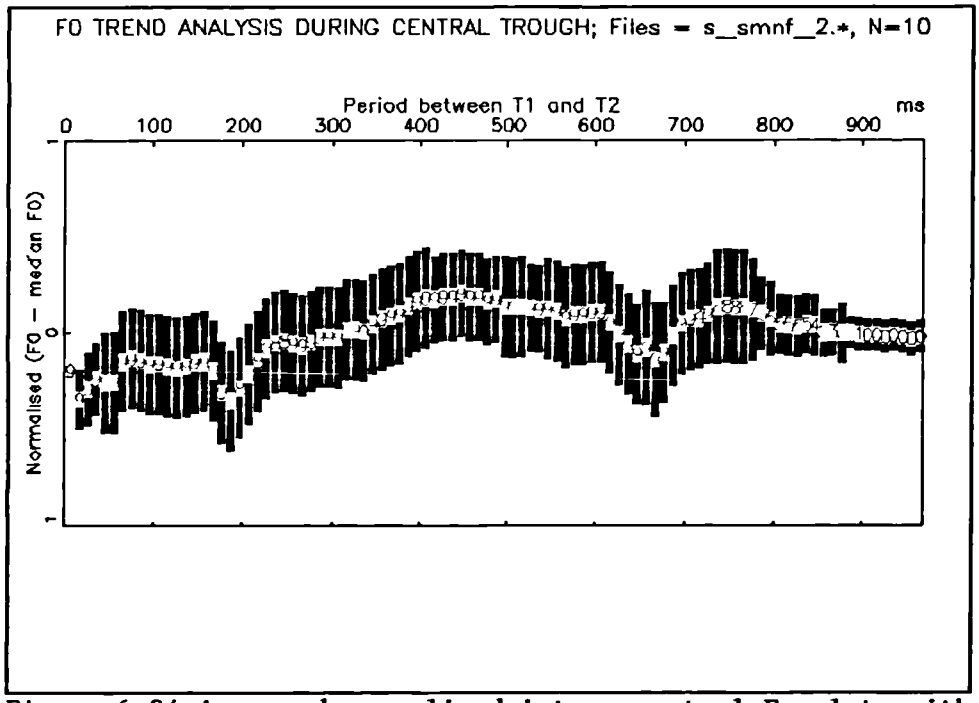


Figure 6.24 Averaged normalised interaccentual Fx plot, with selective masking noise alternating with speech feedback, SPEAKER C. Short sentence data.

CHAPTER 7
CONCLUSION

7.1 INITIAL REMARKS

At the beginning of this thesis, it was mentioned that there were five strands running through it: Descriptive, Phonological, Metatheoretical, Computational Modelling and Experimental. In this conclusion, each of these strands will be discussed in turn, by way of elucidating the findings of the thesis. This discussion will incorporate suggestions for further research, and other observations about the utility of particular modes of research. The object of study will naturally be incorporated into the discussion, where, as proposed at the very beginning of the thesis, it has a natural and deserving place as an object of study in all five of those areas of investigation, for a number of reasons.

7.2 THE DESCRIPTIVE STRAND

In many ways, the description of the phenomena of phonetics (as here is true of the description of declination) is the hardest part of the exercise, because (i) there is sometimes a sense that something has to be made of a particular phenomenon, so that, in the case of declination, it appears not enough to say that intonation contours tend to decline, any particular instance of the phenomenon being displayed, and acting as a sufficient exposition of it; and (ii) there is a danger of carrying too much theoretical baggage into the enquiry, such that claims are made about the nature of real-world data which are implicitly being modelled in terms of (pre)theoretical constructs, such as 'the baseline' and 'the topline'.

This balancing act between preserving an objective view of things and acceding to the use of theoretical constructs has been exemplified in the work here by the opposition between the Local and Global Declination Hypotheses. The Local Declination Hypothesis is naturally part of a more[1] empirical study of phonetic phenomena, to the extent that it describes only phenomena which can be grabbed by the hand, ear or eye. It makes predictions only upon (the

---

[1] We cannot say 'purely empirical' because any phenomenon, qua percept, is imbued with theoretical constructs. To pursue phonetic research, then, some philosophical stance must be taken, notably about the raw data of study, and, at presumably more advanced levels, about the relationship between neurophysiological phenomena and perceptual phenomena.

grosser aspects of) what is physically present in an intonation contour. At the same time, as part of an approach to research into downward trends into intonation, it struggles when it comes to saying anything more than trivial about the form of declination. It can be associated with statements like "if the intonation contour happens to go down for a while at this point, then, so long as the slope of the descent is not too steep, then that can be considered an instance of declination, which may elicit the 'declination effect' on a following accented syllable. If it doesn't go down, then the likelihood of the declination effect taking place is diminished". It doesn't necessarily predict where the intonation contour will go down. It is part as much of description as of theory.

The Global Declination Hypothesis, on the other hand, has no problem in making statements about the form of declination. In fact, it is burning to do so, to make predictions about the shape of the frame of reference in which accentual movements or peaks and troughs are scaled. The Global Declination Hypothesis is part of a well-established theory whose primitives have incorporated an increasing number of abstract components accounting for downward trends in recent years. Thus a tendency has continued, to wrest from areas such as physiology, biology, and other analog sciences the responsibility for accounting for phenomena which once formed a proper part of their domain of study, but which now form part of essentially automata-theoretic accounts of human communicative capability. Thus as description gives way to prediction, so does phonetics to phonology.

7.3 THE PHONOLOGICAL STRAND

The central pivot in modern phonological accounts of downward trends in intonation is the process of downstep (cf. Pierrehumbert and Beckman 1988). A chapter has been devoted to this area of phonological research in this thesis, but the title of that chapter represents the view that is more closely followed here, that downstep is a phenomenon to be observed in intonation patterns, but doesn't have some covert generative role in their determination. The issue is reflective of a number of issues that could be raised within the realm of phonological research in respect of the appropriateness of systems of symbol manipulation to model human behaviour, but they are more appropriately raised in discussion of the metatheoretical strand.

There is one basic aspect of downstep sequences which makes it attractive to incorporate a concept of downstep derived from them into a more general account of intonation within the generative tradition. That is, downstep sequences, as they were first accounted for in the case of certain African Tone languages, are sequences of static tones. There is thus a natural way of assimilating the process of downstep into a model of intonation which is based on the decomposition and generation of intonation contours into and from tonal elements. Once it has been incorporated into that sort of model, it would be inelegant to prevent it constituting part of a productive process in accounting for downscaling of isolated accents in an intonation contour as well as of sequences of such accents.

It seems to the current author that the important thing about downstep is not, however, the fact that it occurs in the presence of and can be viewed as being responsible for downward trends in intonation, but that it occurs in repetitive sequences of tonal configurations. Repetitive sequences are common in intonation, but it is not common for them all to be scaled at the same level, presumably to allow accentuation processes to occur which would otherwise be hampered by the presence of habituation and the lack of differential salience in the F0 signal. They will thus on occasion naturally occur in downward moving sequences, and sometimes in upward moving sequences, thus making upstep and downstep both equivalent processes of "repetition-" or "copy- adjustment". As far as downward trends in intonation are concerned, then, the process of downstep is only productive in allowing for repetition to be modulated in a way which is in keeping with an already operative background trend.

However, it has also been suggested in this thesis that there is something special about the downward-going sequence of step-accents, the classical downstep sequence, and this has something to do with the degree of prominence that can be imparted by an accent in the configuration of a step. As this is limited, prominence must be imparted by movement, in the production of intonation, onto a neighbour (though the neighbours may all have equal pitch prominence). If intonation contours generally tend to start high, then a classical downstepping sequence would naturally result in the process of imparting prominence.

7.4 THE METATHEORETICAL STRAND

Many of the conflicts in intonational research (such as the 'levels' vs. 'configurations' debate) could turn out to be spurious if the appropriate paradigm is instead adopted for conducting particular aspects of the research. The most important division between paradigms is in the decision to pursue research into a competence model of speech and language or into a performance model of the same. This has been emphasised throughout this thesis. The importance of this decision is reflected by the fact that observations about language, and, a fortiori, speech (of which, crucially intonation, being less clearly segmental in structure) made in the one paradigm aren't easily made in that paradigm without being misinterpreted or just plain false if transferred into the other paradigm.

For instance, any debate about the psychological reality of tonal primitives in an intonational model is pretty meaningless unless the full metatheoretical terms of the model in which they are used are expressed. In the sense of 'correspondence to entities manipulated within the human brain or mind', that question only makes sense within a performance model of language. It is one of the tenets of this thesis that static tones could have unique status as psychologically real entities in a performance model of intonation; they just don't happen to have it, though they may share it with pitch movements. Static tones are, though, uniquely serviceable in a competence model of intonation, for which unanalysed primitives are the correct parameters.

Devising a performance model of intonation is not an easy task. The more detail one encompasses in such a model, the more extensive the task, of course, which means that, practically, the model must needs be simplified. However, the initial statement of position in this thesis was that declination was a particularly appropriate area to work in if the pursuit of a performance model of intonation is the target, since its basic pattern is the most rudimentary, and its underlying mechanisms perhaps the most primitive, in the intonation-processing systems of an individual human.

In this regard, it is a natural exercise, in a continuing study of the auditory control of declination, to speculate on the possible physiological mechanisms that could be responsible for it, should it exist in a particular form. This has led the author to spend time reading the neurophysiological literature in an

investigation of the likelihood that a subcortical auditory feedback loop exists for the control of the laryngeal and respiratory musculature, on which the following brief comments can be made.

The level at which collateral projections in a putative feedback loop from nuclei in the ascending auditory pathway would be made to nuclei in the laryngeal and pulmonary motor pathways would depend on the particular reference signal expected as input to the feedback circuitry. For the detection of slope by specialised slope detectors, which is one of the mechanisms it is suggested in this thesis could be used, the author has found no evidence of the existence of appropriate detectors below the cortical level. Langner (1992) reports that, for the cat, Whitfield and Evans (1965) found that 4% of neurons investigated discharged preferentially to ascending or descending F0 contours, which is at least the basic prerequisite for such a kind of detection.

For a general feedback mechanism, perhaps the Inferior Colliculus would be the most appropriate transmitter of the test signal and receiver of the reference signal. It is physically located near to the mid-brain periaqueductal grey (PAG), which it is suggested is used for the coordination of respiratory and phonatory mechanisms (Depaulis and Bandler 1991), and there are known to be collateral projections from the inferior colliculus to PAG (at least in the rabbit - Meller and Dennis, 1986). If high accuracy of periodicity coding were required in a feedback loop (and this would be one way of detecting local slope) this could be provided by the periolivary region[2] in the Superior Olivary Complex, which could perhaps project to levels in the laryngeal motor pathway below PAG (for instance, the nucleus ambiguus).

7.5 THE EXPERIMENTAL STRAND

It is not, of course, appropriate to perform invasive experimentation on human beings to detect the effect of manipulation on such a putative auditory feedback loop. In principle, it is not necessary either; given accurate models of human periodicity and pitch processing, it should be possible to devise experiments in which appropriate manipulation of administered signals elicits

---

[2] I am grateful to Professor Ray Meddis for this suggestion.

behaviour which requires interpretation in a restricted set of ways. This form of experimentation, it seems to the author, is the most likely means of achieving correct results in research into intonational performance (Cf. Siegel et al. 1984).

The experiment reported on in Chapter 6 could be improved on in a number of ways. Firstly, an utterance token could be chosen which allowed for triggering of each accent in the two-peaked accent utterance using criteria other than threshold F0; the onset of a voiceless fricative would provide an appropriate episode for this purpose. Its use would prevent the problem that, because of a built-in minimum latency of 50ms in the buffering mechanism used by the D-to-A control processor, the headphone-administered signal was always delayed by at least this amount (but no more than ten or twenty ms. more than this amount) relative to the triggering event, which, given the rapidity of accent-onset, meant that half the accent was masked when it should have been unmasked.

Secondly, different types of utterance could be elicited in random order, to reduce the possibility of habituation.

Finally, two measures could be taken to attempt to counteract the Lombard Effect. Firstly, a target intonation contour could be chosen which favoured much more strongly a declining interaccentual stretch, such as a falling head configuration. The use of the fall-rise nuclear contour in English could be considered to favour such a context. Secondly, if Lane and Tranel's suggestion that the Lombard Effect is attenuated by the absence of the possibility of meaningful communication is correct, use of a nonsense utterance as the text would perhaps allow the objects of study to be tested for more clearly.

7.6 THE COMPUTER MODELLING STRAND

One of the advantages of the type of experiment performed in this thesis is that it makes predictions about the interaction between perception and production, which interaction underpins all normal human speech communicative ability. However, there is another way of doing this, and this is to model such activity on a computer. The computer model of one aspect of human intonational performance discussed in Chapter 5 of this thesis makes

certain predictions about the interaction between productive and perceptual processes (albeit in a form abstracted form the human perceptual and productive phsyiological mechanisms). This approach could be elaborated on in the future.

7.7 FINAL REMARKS

The question of the form and possibility of auditory control in declination has not been fully resolved in this thesis. However, the thesis does act as a springboard for later research. In particular, it is important to continue to raise questions about the form and operation of F0 declination, which it is to be hoped, could ultimately yield answers as to the question of the ontogenetic and phylogenetic development of intonation contours. With that aim in mind, the study of declination naturally proceeds hand-in-hand with the study of accentuation, which has been the approach in this thesis.

BIBLIOGRAPHY

Adzaku, F.K. and Wyke, B., (1979), 'Innervation of the Subglottalic Mucosa of the Larynx, and its Significance', Folia Phoniatrica, 31, 271-283.

Adzaku, F.K. and Wyke, B., (1982), 'Laryngeal Subglottic mucosal reflexogenic influences on laryngeal muscle activity', Folia Phoniatrica 34, 57-64.

Anderson, M.D., Pierrehumbert, J.B. and Liberman, M.Y. (1984), 'Synthesis by rule of English intonation patterns', Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2.8.1-2.8.4.

Ashby, M.A. (1990), 'Prototype Categories in Phonetics', Speech, Hearing and Language, UCL, pp.19-28.

Atkinson, J.E. (1978), 'Correlation Analysis of the Physiological Factors controlling Fundamental Voice Frequency', JASA 63.1, 211-22.

Baer, T., Sasaki, K. and Harris, K. (eds.) (1987), 'Laryngeal Function in Phonation and Respiration', San Diego, College-Hill Press.

Barry, W.J., Goldsmith, M. and Fourcin, A.J. (1989), 'A Long-term study of voice fundamental frequency', STA final report, UCL.

Beckman, M.E. and Pierrehumbert, J.B. (1986), 'Intonational Structure in Japanese and English', Phonology Yearbook 3, 255-309.

Bland, B.H. (1986), 'The Physiology and Pharmacology of Hippocampal Formation Theta Rhythms', Progress in Neurobiology, 26, 1-54.

Bless, D.M. and Abbs, J.H. (eds.) (1983), 'Vocal Fold Physiology: Contemporary Research and Clinical Issues', San Diego, College-Hill Press.

Bolinger D. (1965), 'Forms of English: Accent, Morpheme, Order'. Tokyo: Hokou; Cambridge: Harvard). Ed. by I. Abe and T. Kanekiyo.

Bolinger, D. (1978), 'Intonation across Languages' in Greenberg, J.H., Ferguson, C.A. and Moravsck, E.A. (eds.), 'Universals of Human Language; Vol. 2, Phonology', 471-524.

Bolinger, D. (1986), 'Intonation and its parts', Edward Arnold.

Bruce, G. (1977), 'Swedish word accents in sentence perspective', Lund: Gleerup.

Chomsky, N. (1965), 'Aspects of the Theory of Syntax', Cambridge (MA), MIT Press.

Chomsky, N. (1988), 'The Generative Enterprise', Foris, Dordrecht.

Chomsky, N. and Halle, M. (1968), 'The Sound Pattern of English', New York, Harper and Row.

Clements, G.N. (1979), 'The Description of Terraced-Level Tone Languages', Language, 55.3, 536-58.

Clements, G.N. (1980), 'The Hierarchical Representation of Tone Features', in Harvard Studies in phonology, Vol. II.

Clements, G.N. and Keyser S.J. (1983), 'CV Phonology', Linguistic Inquiry Monograph 9.

Cohen, A., Collier, R. and 't Hart, J. (1982), 'Declination: Construct or Intrinsic feature of Speech Pitch?', Phonetica 39, 254-73.

Cohen, A. and 't Hart, J. (1967), 'On the anatomy of intonation', Lingua 19: 177-92.

Cole, R.A. (ed.) (1980), 'Production and Perception of Fluent Speech'. Lawrence Erlbaum Associates, Hillsdale.

Collier, R. (1974), 'Laryngeal muscle activity, Subglottal air pressure, and the control of pitch in speech', Haskins Labs. Status Report on Speech Research, SR-39/40.

Collier, R. (1985), 'The setting and resetting of the baseline', Ann. Bull. RILP, No. 19, 111-32.

Collier, R. (1987), 'F0 declination: the control of its setting, resetting and slope', in Baer et al. (eds).

Collier, R. and 't Hart, J. (1981), 'Cursus Nederlandse Intonatie', Leuven/Amersfoort: Acco/De Horstink.

Collier, R. (1989), 'Intonation Analysis: the Perception of Speech melody in relation to Acoustics and Production', Proc. Eurospeech Conf., Paris, Vol. 1, 38-44.

Collier, R. (1990), 'On the Perceptual Analysis of Intonation', Speech Communication, 9, 443-51.

Cooper, W.E. and Sorensen, J.M. (1981), 'Fundamental Frequency in Sentence Production', Berlin, Springer.

Crystal, D. (1969), 'Prosodic Systems and Intonation in English', Cambridge, CUP.

Cutler, A. and Ladd, D.R. (eds.) (1983), 'Prosody: Models and Measurements', Berlin, Springer.

Davis, P.J. and Zhang, S.P. (1991), 'What is the role of the Midbrain Periaqueductal Gray in Respiration and Vocalisation?' in Depaulis, A. and Bandler, R. (eds.).

De Pijper, J. R. (1983), 'Modelling British English Intonation'. Dordrecht: Foris.

Depaulis, A. and Bandler, R. (eds.) (1991), 'The Midbrain Periaqueductal Gray Matter; Functional, Anatomical and Neurochemical Organisation', NATO-ASI Series, New York, Plenum.

Dixon Ward, W. (1970), 'Temporary Threshold Shift and Damage-Risk Criteria for Intermittent Noise Exposures', JASA 48.2/2, 561-74.

Edward, J.A. (1982), 'Rules for Synthesising the Prosodic Features of Speech', JSRU Research Report No. 1015.

Flanagan, J.L. (1972), 'Speech analysis, synthesis and perception', New York, Springer.

Fourcin, A.J. (1962), 'An Aspect of the perception of pitch", Proc. Phon. Sci. IV, 355-9.

Fourcin, A.J. (1974), 'Laryngographic Examination of Vocal Fold vibration', in Wyke, B. (ed.) 'Ventilatory and Phonatory Control Systems: An International Symposium', London, OUP, 315-26.

Fourcin, A.J. and Abberton, E. (1971), 'First applications of a new laryngograph', Med. Biol. Illus. 21, 172-82.

Fry, D.B. (1958), 'Experiments in the perception of stress', Language and Speech, 1, 126-52.

Fudge, E.C. (1969), 'Syllables', Journal of Linguistics 5, 253-86.

Fujimura, O. (ed.) (1988), 'Vocal Fold Physiology: Voice Production, Mechanisms and Functions', New York, Raven Press.

Fujisaki, H. (1987), 'A Note on the Physiological and Physical Basis for the Phrase and Accent Components in the Voice Fundamental Frequency Contour', Ann. Bull. RILP No. 21, 165-75.

Fujisaki, H. and Hirose, K. (1982), 'Modelling the dynamic characteristics of voice fundamental frequency with applications to analysis and synthesis of intonation', in 'Preprints of Papers, Working Group on Intonation, 13th International Congress of Linguists', Tokyo, 57-70.

Gauffin, J. and Hammarberg, B. (eds.) (1991), 'Vocal Fold Physiology: Acoustic, Perceptual and Physiological Aspects of Voice Mechanisms', San Diego, Singular Publishing group Inc.

Grice, M. (1992), 'The Intonation of Interrogation in Palermo Italian; Implications for Intonation Theory', Unpublished PhD Thesis, University College London.

Goldsmith, J.A. (1990), 'Autosegmental and Metrical Phonology', Oxford, Blackwell.

Greenwood, D.D. (1961) 'Critical Bandwidth and the frequency co-ordinates of the basilar membrane', JASA 33, 1344-56.

Gussenhoven, C. and Rietveld, A.C.M. (1988), 'Fundamental Frequency declination in Dutch: testing three hypotheses', Journal of Phonetics 16, 355-69.

Gussenhoven, C. and Rietveld, A.C.M. (1989), 'Reply to Terken', Journal of Phonetics 17, 364->

Halliday, M.A.K. (1967), 'Intonation and Grammar in British English'. The Hague, Mouton.

Halliday, M.A.K. (1970), 'A Course in Spoken English: Intonation'. Oxford University Press, London.

Hamon, C., Moulines, E. and Charpentier, F. (1989), 'A Diphone Synthesis System based on Time-Domain Prosodic Modifications of Speech', Proc. ICASSP, Glasgow, 238-41.

Hermes, D.J. and van Gestel, J.C. (1991), 'The frequency scale of speech intonation', JASA 90.1, 97-102.

Hess, W.J. (1983), 'Pitch Determination of Speech Signals - Algorithms and Devices', Berlin, Springer.

't Hart, J, Collier, R. and Cohen, A. (1990), 'A Perceptual Study of Intonation'. Cambridge Studies in Speech Science and Communication, Cambridge University Press.

't Hart, J. (1986), 'Declination has not been defeated', JASA 80.6, 1839-40.

Hewitt, M.J. and Meddis, R. (1993), 'Regularity of Cohlear Nucleus Stellate Cells: A Computational Modelling Study', JASA 93.6, 3390-99.

Hewitt, M.J. and Meddis, R. (forthcoming), 'A Computer Model of Amplitude-Modulation Sensitivity of Single Units in the Inferior Colliculus', Draft submitted to JASA.

Hewitt, M.J., Meddis, R. and Shackleton, T.M. (1992), 'A Computer Model of a Cochlear-Nucleus Stellate Cell: Responses to Amplitude-modulated and Pure-tone Stimuli', JASA 91.4, 2096-2109.

Hirst, D.J. (1983), 'Structure and Categories in Prosodic Representations', in Cutler and Ladd (eds.).

Hirst, D.J. (1988), 'Tonal Units as Phonological Constituents: the evidence from French and English Intonation', in Van der Hulst, H. and Smith, N. (eds.) 'Autosegmental Studies in Pitch Accent', Foris, Dordrecht.

Hirst, D.J. and Espesser, R. (1991), 'Automatic modelling of Fundamental Frequency', Travaux de l'Institut de Phonetique, 15, Aix en Provence.

Hjelmslev, L. (1953), 'Prolegomena to a Theory of Language', Bloomington, Indiana University.

Holmes, J.N., Mattingly, I.G. and Shearme, J.N. (1964), 'Speech Synthesis by Rule', Language and Speech 7, 127-43.

House, J. (1989), 'Syllable structure constraints on F0 timing', Poster presented at Lab. Phon. 2 conference, Edinburgh.

House, J. and Johnson, M, (1987), 'Enlivening the Intonation in Text-to-Speech Synthesis: An "Accent-Unit" Model', Proc. XIth Int. Cong. of Phonetic Sciences, Tallinn, Estonia.

House, J., Johnson, M. (1989), 'Improvements to Speech Synthesis-by-rule Algorithms; Summary and Technical Report', MOD research document, UCL.

House, J. and Youd, N. (1991), 'Stylised Prosody in Telephone Information Services: Implications for Synthesis', Proc. XII I.C.Ph.S., 198-201.

Huss, V. (1978), 'English word stress in the post-nuclear position', Phonetica 35, 86-105.

Jakobson, R., Fant, G. and Halle, M. (1952), 'Preliminaries to Speech Analysis: the Distinctive Features and their Correlates', Cambridge (MA), MIT Press.

Jassem, W. (1952), 'The Intonation of Conversational English', Travaux de la Societe des Sciences et Lettres, Wroclaw.

Johnson, M.E. (1990), 'Implementation of an intonation algorithm for synthesis by rule', Speech Hearing and Language, UCL, 195-226.

Johnson, M. and Grice, M. (1990a), 'The Phonological Status of Stylised Intonation Contours', Speech, Hearing and Language, UCL, Vol.4.

Johnson, M. amd Grice, M. (1990b), 'Some Phonetic Correlates of Stylisation in the Step-Down Contour', Proc. IOA Autumn Conference on Speech and Hearing.

Johnson, M. and House, J. (1986), 'An accent-unit model of intonation for text-to-speech sythesis', Proc. IOA Conf. on Speech and Hearing 8, 409-16.

Juergens, U. and Pratt, R. (1979a), 'Role of the Periaqueductal Grey in Vocal Expression of Emotion', Brain Research, 167, 367-78.

Juergens, U. and Pratt, R. (1979b), 'The Cingular Vocalization Pathway in the Squirrel Monkey', Experimental Brain Research, 34, 559-70.

Kingdon, R. (1958), 'The Groundwork of English Intonation', Longman.

Kraayefeld, H. (1992), Personal communication.

Kubozono, H. (1989), 'Syntactic and rhythmic effects on downstep in Japanese', Phonology 6.1, 39-67.

Ladd, D.R. (1978), 'Stylised Intonation', Language 54, 517-41.

Ladd, D.R. (1980), 'The Structure of Intonational meaning: Evidence from English'. Bloomington: Indiana University Press.

Ladd, D.R. (1983), 'Phonological Features of Intonational Peaks', Language, 59.4, 721-59.

Ladd, D.R. (1984), 'Declination: a review and some hypotheses', Phonology 1, 53-75.

Ladd, D.R. (1988), 'Declination "reset" and the hierarchical organisation of utterances', JASA 84, 530-44.

Ladd, D.R. (1990), 'Metrical Representation of Pitch Register', in Kingston and Beckman (eds.) 'Papers in Laboratory Phonology', CUP Press.

Ladd, D.R. (1992), 'Introduction to Intonational Phonology', in Dougherty, G. and Ladd, D.R. (eds.) 'Papers in Laboratory Phonology II, Gesture, Segment, Prosody, Cambridge, CUP.

Ladd, D.R. and Johnson, C. (1987), ' "Metrical" factors in the scaling of sentence-initial accent peaks', Phonetica 44, 238-45.

Ladefoged, P. (1963), 'Some physiological parameters in speech', Language and Speech 6, 1963.

Ladefoged, P. (1971), 'Preliminaries to Linguistic Phonetics', Chicago University Press.

Lane, H. and Tranel, B. (1971), 'The Lombard Sign and the Role of Hearing in Speech', Journal of Speech and Hearing Research 14, 677-709.

Langner, G. (1992), 'Periodicity Coding in the Auditory System', Hearing Research 60, 115-42.

Leroy, L. (1984), 'The Psychological Reality of Fundamental Frequency Declination', Antwerp Papers in Linguistics, Nr. 40.

Liberman, M.Y. (1975) 'The Intonation System of English. PhD Thesis, MIT.

Liberman, M. and Sag, I.A. (1974), 'Prosodic Form and Discourse Function', in Papers from 10th. Regional Meeting, Chicago Linguistic Society, 416-427.

Liberman, M. and Prince, A. (1977), 'On stress and linguistic rhythm', Linguistic Inquiry 8, 249-336.

Liberman, M. and Pierrehumbert, J.B. (1984), 'Intonational Invariance under changes in pitch range and length', in 'Language Sound Structure', Aronoff, M. and Oehrle, R. (eds.), Cambridge (MA), MIT Press, 157-233.

Liberman, A.M., Cooper, F.S., Shankweiler, D.S. and Studdert-Kennedy, M. (1967), 'Perception of the Speech Code', Psychol. Review 74, 431-461.

Liberman, A.M. and I.G. Mattingly (1985), 'The motor theory of speech perception revised', Cognition 21, 1-36.

Lieberman, P., Katz, W., Jongman, A, Zimmerman, R. and Miller, M. (1985), 'Measures of the sentence intonation of read and spontaneous speech in American English', JASA 77, 649-57.

Maeda, S. (1976), 'A characterization of American English Intonation', Unpublished PhD dissertation, MIT.

Marcus, S.M. (1976), 'Perceptual Centres', PhD Thesis, University of Cambridge.

Mattingly, I.G. (1966), 'Synthesis by rule of prosodic features', Language and Speech 9, 1-15.

McAllister (1971), quoted in Cooper and Sorensen (1981).

Mead, K.O. (1974), 'Identification of Speakers from Fundamental Frequency Contours in Conversational Speech', JSRU Report 1002.

Meddis, R. and Hewitt, M.J. (1991a), 'Virtual Pitch and Phase Sensitivity of a Computer Model of the Auditory Periphery. I: Pitch Identification', JASA 89.6, 2866-82.

Meddis, R. and Hewitt, M.J. (1991b), 'Virtual Pitch and Phase Sensitivity of a Computer Model of the Auditory Periphery. II: Phase Sensitivity', JASA 89.6, 2883-94.

Meller, S.T. and Dennis, B.J. (1986), 'Afferent Projections to the Periaqueductal Gray in the Rabbit', Neuroscience, 19.3, 927-64.

Mosteller, F. and Tukey, J. (1977), 'Data analysis and regression', Reading (MA), Addison-Wesley.

Nolan, F. (1984), 'Auditory and instrumental analysis of intonation', Cambridge Papers in Phonetics and Experimental Linguistics 3, Dept. of Linguistics, University of Cambridge.

O'Connor, J.D. and Arnold, G.F. (1973), 'Intonation of Colloquial English', London, Longman.

Ohala, J.J. and Jaeger, J. (1986), 'Experimental Phonology'.

Ohala, J.J. (1970), 'Aspects of the Control and Production of Speech', UCLA Working Papers in Phonetics, 15.

Ohala, J.J. (1982), 'Physiological Mechanisms underlying Tone and Intonation', Working Group on Intonation, Fujisaki, H. and Garding, E. (orgs.), XIIIth Int. Congress of Linguists, Tokyo.

Ohala, J.J. (1984), 'An Ethological Perspective on Common Cross-Language Utilisation of F0 of Voice', Phonetica 41, 1-16.

Ohman, S. (1967), 'Word and Sentence Intonation: A quantitative model', STL-QPSR 2-3, 20-54.

O'Shaughnessy, D. (1976), 'Modelling Fundamental Frequency and its Relationship to Syntax, Semantics and Phonetics', PhD Thesis, MIT.

Pandit, S.M. and Wu, S.M. (1983), 'Time Series and System Analysis with Applications', New York, Wiley.

Pierrehumbert, J. (1979), 'The Perception of Fundamental Frequency Declination', JASA 66, 363-9.

Pierrehumbert, J.B. (1980), 'The Phonology and Phonetics of English Intonation', MIT dissertation.

Pierrehumbert, J. (1981), 'Synthesizing Intonation', JASA 70.4, 985-95.

Pierrehumbert, J.B. and Beckman, M.E. (1988), 'Japanese Tone Structure', Linguistic Inquiry Monographs:15, Cambridge (MA), MIT Press.

Pierrehumbert, J.B. and Steele, S.A. (1987), 'How many Rise-Fall-Rise Contours?', Proc. XIth. Intl. Congress of Phonetic Sciences, Tallinn, Estonia. Se49.1 145-8.

Pike, K.L. (1945), 'The Intonation of American English', University of Michigan press, Ann Arbor.

Poeppel, E. and Logothetis, N. (1986), 'Neuronal Oscillations in the Human Brain', Naturwissenschaften, 73, 267-8.

Popper, K.R. (1957), 'The aim of Science', Ratio 1, 24-35.

Repp, B.H. (1985), 'Critique of "Measures of the sentence intonation of read and spontaneous speech in American English"[J.Acoust.Soc.Am.77,649-657(1985)]', JASA 78.3, 1114-5.

Scheffe, H. (1952), 'An Analysis of Variance for Paired Comparisons' Journal of the American Statistical Association, 47, 381-400.

Scuffil, M. (1982), 'Experiments in Comparative Intonation', Max Niemeyer Verlag, Tubingen.

Selkirk, E.O. (1984), 'Phonology and Syntax', Cambridge (MA), MIT Press.

Siegel, G.M., Pick, Jr. H.L. and Garber, S.R. (1984), 'Auditory Feedback and Speech Development', Advances in Child Development and Behaviour 18, 49-79.

Silverman, K.E.A. and Pierrehumbert, J.B. (1990), 'The timing of Prenuclear High Accents in English', in Kingston, J. and Beckman, M.E., Papers in Laboratory Phonology, Cambridge, CUP.

Silverman, K.E.A. (1987), 'The Structure and Processing of Fundamental Frequency Contours', Unpublished PhD thesis, University of Cambridge.

Stevens, K.N., Hirano, M. and Abbs, J.H. (eds.) (1981), 'Vocal Fold Physiology', Tokyo, University of Tokyo Press.

Sugito, M. and Hirose, H. (1988), 'Production and Perception of Accented Devoiced Vowels in Japanese', Ann. Bull. RILP, No. 22, 21-39.

Terhardt, E. (1974), 'Pitch, Consonance, and Harmony', JASA 55.5, 1061-9.

Terhardt, E., Stoll, G. and Seewann, M. (1982), 'Algorithm for Extraction of Pitch and Pitch Salience from Complex Tonal Signals', JASA 71.3, 679-88.

Terken, J.M.B. (1989a), 'Reaction to C. Gussenhoven and A.C.M. Rietveld: "Fundamental frequency declination in Dutch: testing three hypotheses:" ', Journal of Phonetics 17, 357-64.

Terken, J.M.B. (1989b), 'Fundamental Frequency and Perceived Prominence of Accented Syllables', IPO Ann. Prog. Report 24, 33-42.

Terken, J.M.B. (1991), 'Fundamental Frequency and perceived prominence of accented syllables', JASA 89(4), 1768-76.

Thorsen, N. (1978), 'An Acoustical Analysis of Danish Intonation', J. Phon. 6, 151-75.

Thorsen, N. (1979), 'Interpreting Raw Fundamental Frequency tracings of Danish', Phonetica 36, 57-8.

Thorsen, N. (1980), 'A Study of the Perception of Sentence Intonation - Evidence from Danish', JASA 67.3, 1014-30.

Thorsen, N. (1983), 'Two Issues in the Prosody of Standard Danish', in Cutler and Ladd (eds.), 27-38.

Thorsen, N. (1985), 'Intonation and text in Standard Danish', JASA 77.3, 1205-16.

Thorsen, N. (1986), 'Sentence Intonation in Textual Context - Supplementary Data', JASA 80.4, 1041-7.

Titze, I.R. (1973), 'The Human Vocal Cords: A Mathematical Model, Part I', Phonetica 28, 129-70.

Titze, I.R. (1974), 'The Human Vocal Cords: A Mathematical Model, Part II', Phonetica 29, 1-21.

Titze, I.R. (1989), 'On the relation between subglottal pressure and fundamental frequency in phonation', JASA 85.2, 901-6.

Titze, I.R. and Durham, P. (1987), 'Passive mechanisms influencing fundamental frequency control', in Baer et al. (eds.).

Titze, I.R. and Scherer, R.C. (eds.) (1983), 'Vocal Fold Physiology: Biomechanics, Acoustics, and Phonatory Control', Denver, The Denver Center for the Performing Arts.

Titze, I.R. and Talkin, D.T. (1979), 'A theoretical study of the effects of various laryngeal configurations on the acoustics of phonation', JASA 66.1, 60-74.

Tonndorf, J. (1976), 'Bone Conduction', in Keidel, W.D. and Neff, W.D., (eds.) 'Handbook of Sensory Physiology', Vol. 3, Berlin, Springer, 37-84.

Trager, G.L. and Smith, H.L. (1951), 'An Outline of English Structure'. Battenberg, Norman, Oklahoma.

Umeda, N. (1982), ' "F0 declination" is situation dependent', Journal of Phonetics 10, 279-90.

Vayra, M. and Fowler, C.A. (1992), 'Declination of Supralaryngeal Gestures in Spoken Italian', Phonetica 49, 48-60.

Vertes, R.P. (1982), 'Brain Stem Generation of the Hippocampal EEG', Progress in Neurobiology 10, 159-86.

Vinogradova, O.S. (1976), 'Functional Organisation of Limbic System' in Isaacson, R.L. and Pribram, K.H. (eds.) 'The Hippocampus; Vol 2, Neurophysiology and Behaviour', New York, Plenum Press.

Whitfield, I.C. and Evans, E.F. (1965), 'Responses of auditory cortical neurons to stimuli of changing frequency', Journal of Neurophysiology 28, 655-72.

Wyke, B.D. (1974), 'Laryngeal Myotatic reflexes and phonation', Folia Phoniatrica 26, 249-64.

ADDENDUM

p.145 Line 18: insert footnote just after "subtree", as follows: "In the performance model of local accentuation and declination presented in Chapter 5, a similar use is made of both a lookback and lookahead mechanism. However, the mechanisms are restricted to the domain of two accents, except in the case of a downstepped accent, in which there is global lookahead. This is contrary to the situation in Ladd's competence model."